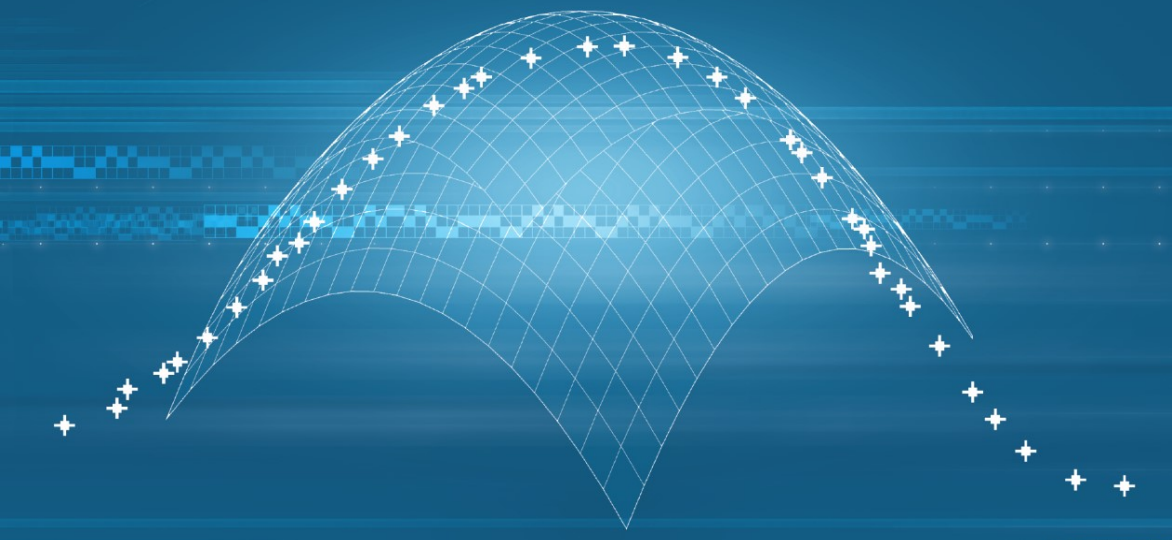
INTERNATIONAL SERIES IN OPERATIONS RESEARCH AND MANAGEMENT SCIENCE

Process Optimization

A Statistical Approach



Enrique del Castillo

Springer

Process Optimization
A Statistical Approach

Recent titles in the
INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE
Frederick S. Hillier, Series Editor, *Stanford University*

Gass & Assad/ *AN ANNOTATED TIMELINE OF OPERATIONS RESEARCH: An Informal History*

Greenberg/ *TUTORIALS ON EMERGING METHODOLOGIES AND APPLICATIONS IN OPERATIONS RESEARCH*

Weber/ *UNCERTAINTY IN THE ELECTRIC POWER INDUSTRY: Methods and Models for Decision Support*

Figueira, Greco & Ehrgott/ *MULTIPLE CRITERIA DECISION ANALYSIS: State of the Art Surveys*

Reveliotis/ *REAL-TIME MANAGEMENT OF RESOURCE ALLOCATIONS SYSTEMS: A Discrete Event Systems Approach*

Kall & Mayer/ *STOCHASTIC LINEAR PROGRAMMING: Models, Theory, and Computation*

Sethi, Yan & Zhang/ *INVENTORY AND SUPPLY CHAIN MANAGEMENT WITH FORECAST UPDATES*

Cox/ *QUANTITATIVE HEALTH RISK ANALYSIS METHODS: Modeling the Human Health Impacts of Antibiotics Used in Food Animals*

Ching & Ng/*MARKOV CHAINS: Models, Algorithms and Applications*

Li & Sun/ *NONLINEAR INTEGER PROGRAMMING*

Kaliszewski/ *SOFT COMPUTING FOR COMPLEX MULTIPLE CRITERIA DECISION MAKING*

Bouyssou et al/ *EVALUATION AND DECISION MODELS WITH MULTIPLE CRITERIA: Stepping stones for the analyst*

Blecker & Friedrich/ *MASS CUSTOMIZATION: Challenges and Solutions*

Appa, Pitsoulis & Williams/ *HANDBOOK ON MODELLING FOR DISCRETE OPTIMIZATION*

Herrmann/ *HANDBOOK OF PRODUCTION SCHEDULING*

Axsäter/ *INVENTORY CONTROL, 2nd Ed.*

Hall/ *PATIENT FLOW: Reducing Delay in Healthcare Delivery*

Józefowska & Węglarz/ *PERSPECTIVES IN MODERN PROJECT SCHEDULING*

Tian & Zhang/ *VACATION QUEUEING MODELS: Theory and Applications*

Yan, Yin & Zhang/ *STOCHASTIC PROCESSES, OPTIMIZATION, AND CONTROL THEORY APPLICATIONS IN FINANCIAL ENGINEERING, QUEUEING NETWORKS, AND MANUFACTURING SYSTEMS*

Saaty & Vargas/ *DECISION MAKING WITH THE ANALYTIC NETWORK PROCESS: Economic, Political, Social & Technological Applications w. Benefits, Opportunities, Costs & Risks*

Yu/ *TECHNOLOGY PORTFOLIO PLANNING AND MANAGEMENT: Practical Concepts and Tools*

Kandiller/ *PRINCIPLES OF MATHEMATICS IN OPERATIONS RESEARCH*

Lee & Lee/ *BUILDING SUPPLY CHAIN EXCELLENCE IN EMERGING ECONOMIES*

Weintraub/ *MANAGEMENT OF NATURAL RESOURCES: A Handbook of Operations Research Models, Algorithms, and Implementations*

Hooker/ *INTEGRATED METHODS FOR OPTIMIZATION*

Dawande et al/ *THROUGHPUT OPTIMIZATION IN ROBOTIC CELLS*

Friesz/ *NETWORK SCIENCE, NONLINEAR SCIENCE AND DYNAMIC GAME THEORY APPLIED TO THE STUDY OF INFRASTRUCTURE SYSTEMS*

Cai, Sha & Wong/ *TIME-VARYING NETWORK OPTIMIZATION*

Mamon & Elliott/ *HIDDEN MARKOV MODELS IN FINANCE*

**** A list of the early publications in the series is at the end of the book ****

PROCESS OPTIMIZATION
A Statistical Approach

ENRIQUE DEL CASTILLO
Professor, Department of Industrial & Manufacturing Engineering and Department of Statistics, Pennsylvania State University

Springer

Enrique del Castillo
The Pennsylvania State University
Pennsylvania
USA

Library of Congress Control Number: 2007922933

ISBN 978-0-387-71434-9 (HB) ISBN 978-0-387-71435-6 (e-book)

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Para Mónica, Fernando y
Quique.

Contents

Preface

This book is intended as a textbook for a second course in experimental optimization techniques for industrial production processes and other "noisy" systems where the main emphasis is process optimization. This includes courses in "Response Surface Methods" and related topics. The book has outgrown from class notes of a graduate course that I have given for the past 10 years to Industrial Engineering and Operations Research students at Penn State University and at the University of Texas at Arlington. Typically, students come to this course with some background in either Design of Experiments (DOE) or Linear Regression. Many students also come to the course with a background in optimization methods. After teaching this course for several years based on other DOE and Response Surface Methods (RSM) books, it became clear the need for a book more suited to graduate engineering students, who learn about a wide variety of optimization techniques in other courses yet are somewhat disenchanted because there is no apparent connection between those optimization techniques and DOE/RSM.

The point of view of the book is to provide in the form of a text a contemporary account not only of the classical techniques and tools used in DOE and RSM but also to present relatively more advanced process optimization techniques from the recent literature which, perhaps due to lack of exposure or due to their young age, have not been used that much in industrial practice. The book contains a mix of technical and practical sections, appropriate for a first year graduate text in the subject or useful for self-study or reference.

For a person with a more traditional Statistics or Quality Engineering background, the present book will serve as a reference to techniques that

complement and extend basic process optimization techniques from DOE and RSM, including statistical issues that arise in process optimization, Bayesian methods for process optimization, and an introduction to Stochastic Approximation, Kriging methods and "computer experiment" techniques. For a person with an Operations Research background which includes mathematical programming techniques, the present book will not only serve as a guide to DOE and RSM, but will show how important statistical considerations need to be taken into account while optimizing a noisy process.

The book contents are as follows. After an introduction presented in Chapter 1, classical DOE and RSM topics are covered in Part II (Chapters 2 to 5). This includes DOEs for first and 2nd order models (including the concepts of D and G-optimality and an introduction to mixture models), and optimization of these models using classical RSM tools such as steepest ascent, "canonical" analysis, and "Ridge" analysis.

Part III (Chapters 6 to 8) treats the very important issue of sampling variability in an optimization problem where experimental data can vary randomly. Although considerable work has appeared on this subject in the last 15 years in the Statistics literature, this has found very little impact in applications. The effect of sampling variability in the steepest ascent procedure, on the location of the optimal settings, and on the eigenvalues of a quadratic model is discussed. Recent computational techniques for finding confidence regions on the location of the optimal operating conditions of a process are presented here. A discussion of the debate that evolved among opposing schools in design of experiments is presented here (the "bias vs. variance" debate).

Part IV (Chapters 9 and 10) discusses optimization methods that achieve solutions that are insensitive, or robust, to variations in factors that are not controllable. This is the so-called Robust Parameter Design (RPD) problem. A discussion of Split Plot design problems, where there are hard to vary factors in an experiment is included here, as RPD problems often have hard to vary factors. The idea of finding solutions that are not sensitive to variations in the assumed model has become a popular topic in the area of mathematical programming. An approach that uses ideas from robust optimization to solve optimization problems based on models fitted from data is presented in Chapter 10. It is shown that robust optimization methods are closely related to methods

for finding a confidence region in the optimal settings discussed in Chapter 7. Relation of these methods with stochastic programming problems is noted.

Parts I-IV contain what could be described as frequentist statistical techniques for process optimization. In contrast, Part V (Chapters 11 and 12) present recently developed Bayesian methods for process optimization. This is an extremely useful and unfortunately not well-known method for process optimization that resolves many of the issues regarding optimization of a multiple response process based on models fitted from data. First, an overview of Bayesian inference is presented. Then, optimization of single and multiple response systems from a Bayesian perspective is discussed based on the Bayesian multivariate regression model. These approaches provide a means for handling not only the uncertainty in the parameters of the assumed model, but the uncertainty in the model form itself. A Bayesian technique due to Gilmour and Mead for stopping a sequence of experiments based on the gains that are expected from running further experiments is presented. Finally, Bayesian Mixture optimization problems are also discussed. Matlab programs that implement most of the techniques discussed in this chapter are presented.

Part VI deals with design, modeling, and optimization techniques that have received a great deal of interest in recent years but that lie outside the mainstream of techniques usually considered within DOE and RSM. This includes computer experiments, a field with assumptions usually in strong contrast with classic RSM. An introduction to space filling designs and Kriging methods for computer experiments is provided. This part of the book also discusses recently developed stochastic optimization techniques based on stochastic approximation, in particular, Spall's simultaneous perturbation methods. An approach for testing the Karush-Khun-Tucker (KKT) optimality conditions in a problem where models are fitted to noisy responses is presented. Although the methodology presented was developed for optimization of simulated processes, this problem also occurs when optimizing a real or physical (i.e., non-simulated) process.

The book concludes with four appendices on the basics of linear regression, analysis of variance, matrices and optimality results, and statistical results used in Part V. Sections that contain material at a relatively more advanced level are labeled with a **. To facilitate reading, all examples and proofs of theorems end with a ■. MATLAB and MAPLE programs that implement some of the

techniques discussed in the book and solutions to end of chapter problems will be posted on the author's personal web page.

As the Mexican writer Carlos Monsivais has said, one only realizes what one thinks in a topic until it sees it written down. I think such writing activity is very healthy in particular for a university faculty member as myself. Therefore, I would like to thank first of all Professor Fred Hillier for inviting me writing this book. I thank my graduate students who have endured the classes and the preliminary version of the notes on which this book is based. I wish to thank my coauthors and Ph.D. students in RSM topics in recent years: Dr. John J. Peterson (Glaxo SmithKline Beecham), Dr. Ramkumar Rajagopal (formerly at Intel Corp.), Dr. Guillermo Miró-Quesada (Eli-Lilly Co.), Dr. Suntara Cahya (Eli Lilly), Dr. John Semple (Southern Methodist University), and Professors Bert Bentovil and Jack Kleijnen (Tilburg University, The Netherlands). Professor Kleijnen was very kind in reading the chapters on Bayesian optimization, written while he acted as host during my sabbatical visit to his department. I also wish to thank Dr. R. Jean Ruth and Ms. Sharon Zielinski (General Motors R&D), Dr. Mani Janakiram (Intel Corp.), and Dr. Arnon Hurwitz (formerly at SEMATECH) with whom I have had the privilege of working in research related to process modeling and optimization funded by their organizations. I am also indebted to the National Science Foundation, who has been a major source of funding in the last 10 years.

ENRIQUE DEL CASTILLO, STATE COLLEGE, JANUARY 2007

PART I

PRELIMINARIES

Chapter 1

AN OVERVIEW OF EMPIRICAL PROCESS OPTIMIZATION

A masterpiece is an experiment that succeeded.

—Carlos Chávez [1]

1.1 Introduction

Every engineering system or process is designed with an intended purpose. The purpose frequently entails a desired performance of the operation of the product being manufactured or of the process that manufactures it[2]. In many cases, engineering design activities involve tests or experimentation, since the product or process is not well understood, and the desired performance can not be guaranteed. Classical examples abound in Chemical Engineering, in which results from a pilot plant experiment are scaled up to the manufacturing site. In traditional discrete part manufacturing, e.g., machining processes, experimental design and analysis has been used to improve the performance of processes given the inherent noise in the various responses of interest. In designing new products, research & development groups run experiments, build models, and try to optimize responses related to the performance of the new product being designed. In this chapter, we provide an overview of these methods and introduce some basic terminology that will be used in later chapters.

[1] Mexican composer (1899–1978).

[2] We will use the generic word *process* to mean a real manufacturing process of a product, although we will discuss also optimization techniques for simulation and computer *models* of engineering systems (see Part VI).

An interesting high-tech example comes from the semiconductor sector, where expensive manufacturing equipment is seldom utilized "out of the box" as suggested by the equipment manufacturer. Instead, semiconductor manufacturers need to perform a series of experiments in order to achieve a desired performance, in a step in the life of a process called "qualification" in semiconductor circles. Once a process has been qualified, it is "released to manufacturing", which means there is a recommended (hopefully optimal) *recipe* of process settings that should be used in that equipment. Once the process is released to manufacturing, a different set of techniques, mainly feedback control and process monitoring techniques, are implemented to keep the process at the recommended or optimized performance in the presence of further process noise and disturbances (this interplay between optimization and control will be a subject to which we return at different points in this book). Figure 1.1 shows the relation between process optimization and control in this example.

Other cases in engineering where optimization is necessary do not involve a real system, but a model of a real system. This is the case of optimization techniques of simulated systems (e.g., discrete-event simulations) and of some computer codes (e.g., finite element programs) that are very expensive to run.

Some terminology and preliminary concepts are now in order. We will consider a process as a black box. This means that the purpose of a process model is to describe the behavior of the process only based on the observable inputs and outputs. In process optimization and Design of Experiments (**DOE**), the inputs correspond to variables we can control during an experiment, and we will refer to these as **controllable factors**[3]. The outputs of the process correspond to variables we wish to modify by varying the controllable factors, and we will refer to them as the process **responses** and, if the discussion relates to quality control applications, to **quality characteristics**.

The type of processes we consider are "noisy" in the sense that identical settings of the controllable factors will *not* result in identical values of the responses. The variability in the responses is due to other factors not accounted for in the experiment. In classical statistical terminology, we say that

[3]Sometimes, particularly in the literature on Taguchi methods, controllable factors are referred to as parameters, but we will avoid this use.

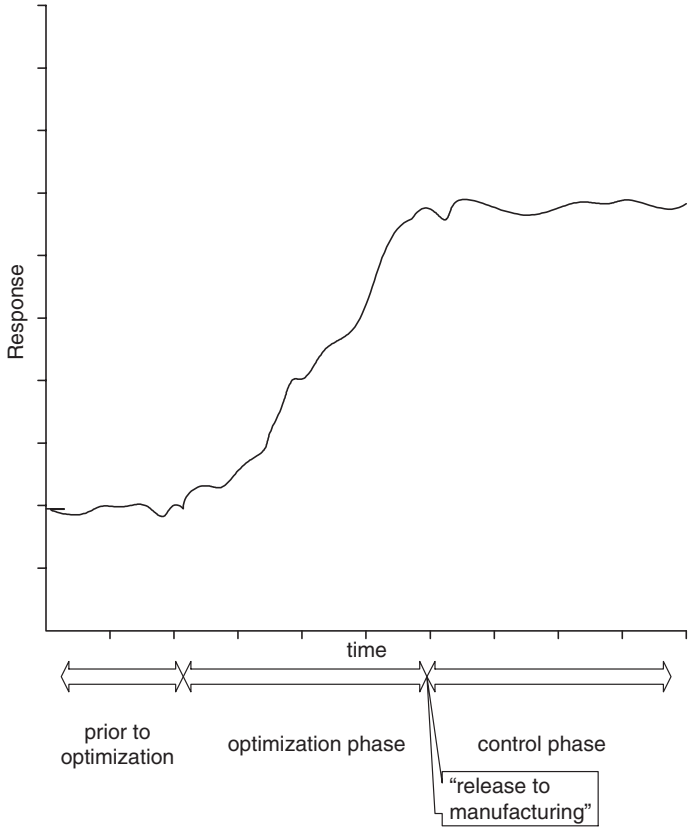


Figure 1.1. The interplay between process optimization and process control in semiconductor manufacturing

the process exhibits sampling variability[4]. It will be assumed the process can be "run" in such a way that steady-state response values can be measured after each "**run**". An **experiment** will then be understood as a set of such experimental runs or tests. The tests are performed on material that is processed; in the DOE literature units of such material are called the **experimental units**. Two typical examples of experimental units in industrial experiments are parts and batches of product. A designed experiment is an experiment in which the tests are planned out before hand. The set of test values for the controllable factors in such plan constitute the **experimental design**.

[4]A different interpretation of this variability is presented in Part V where an introduction to Bayesian Statistics and its use in process optimization is presented.

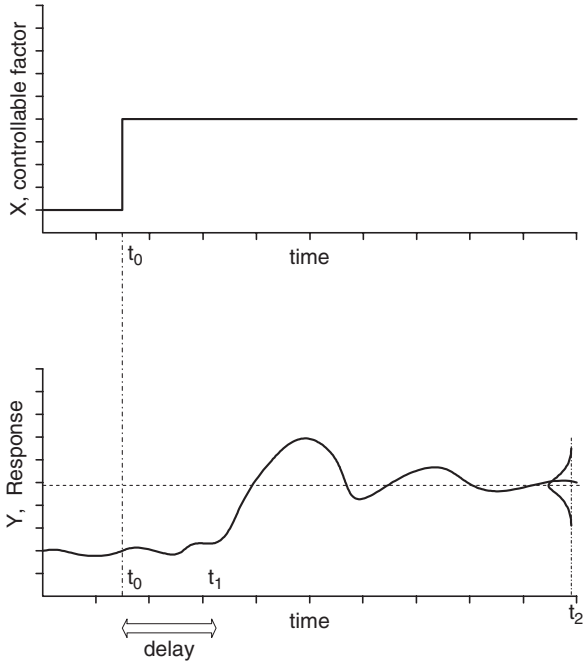


Figure 1.2. Varying a control factor and measuring the steady-state response for a process with inertia

Figure 1.2 shows a response as it varies in a process with inertial elements, such as a chemical process, when a single factor is varied. The controllable factor is varied at time t_0, but the process output starts to vary at time $t_1 > t_0$. The time difference $t_1 - t_0$ is the system delay. After time t_1, the response varies for a while according to some transient response after which the steady state response, i.e., the level of the response that is time-invariant, is eventually obtained. It will be assumed in this book that the process is such that a steady state response is achievable (i.e., the process is "stable" in this sense) and that the interest is in optimizing such steady-state response. Thus, we are willing to wait until time t_2 in the figure to collect measurements. The small vertical normal distribution at time t_2 illustrates that the steady-state value will be subject to sampling variability, so that the steady state response corresponds to the expected value of the distribution of observed responses at $t \geq t_2$. For systems without inertia (i.e., without "process dynamics") it is immediately true that steady state is achievable and of interest, as no transient occurs. Here we have

again the relation with control theory, which typically focuses in optimizing the transient response of a dynamic process.

The Japanese engineer and consultant G. Taguchi [149] introduced the important notion of **noise factors** in the literature on Designed Experiments. These are factors that are not really controllable while the process operates (e.g., environmental temperature or humidity in a typical machining process) or when the final product is used in the marketplace (e.g. driving habits). However, one key insight by Taguchi was to note that these factors can in many cases be controlled (varied) during a carefully conducted experiment. This allows the development of a model in which the responses are a function of both the controllable factors and the noise factors. The purpose of the Taguchi-coined term **Robust Parameter Design** consists of using this model to optimize the controllable factors in such a way that the solution is insensitive or robust with respect to variations in the noise factors. Robust Parameter Design is discussed in Chapter 9.

We will use the following general notation. We assume we have k controllable variables which will be denoted by x_i where $i = 1, 2, \ldots, k$, or, in vector form, as $\boldsymbol{x}$. There may be q noise factors denoted z_i, $i = 1, 2, \ldots, q$ or, in vector form, $\boldsymbol{z}$. There might be m responses of interest which will be denoted by y_i, $i = 1, 2, \ldots, m$, or, in vector form, $\boldsymbol{y}$. Figure 1.3 illustrates these definitions. If the noise factors are uncontrolled in an experiment (i.e., they are "given" by the environment, as in the figure) they are typically not considered in the model, although if they can be measured they could be incorporated in the response model to improve prediction. For Robust Parameter Design, the "environment" in the figure is substituted for another DOE, and the noise factors are then always included in the model.

1.2 Brief Historical Account of Some Process Optimization Methods for Noisy Processes

A short historical sketch of the development of some of the optimization techniques for noisy processes discussed in this book is useful to understand the issues we would like to address later on.

Response Surface Methodology (**RSM**) is a series of experimental design, analysis, and optimization techniques that originated in the work by Box and Wilson in 1951 [29]. The main idea of RSM is to optimize an unknown and

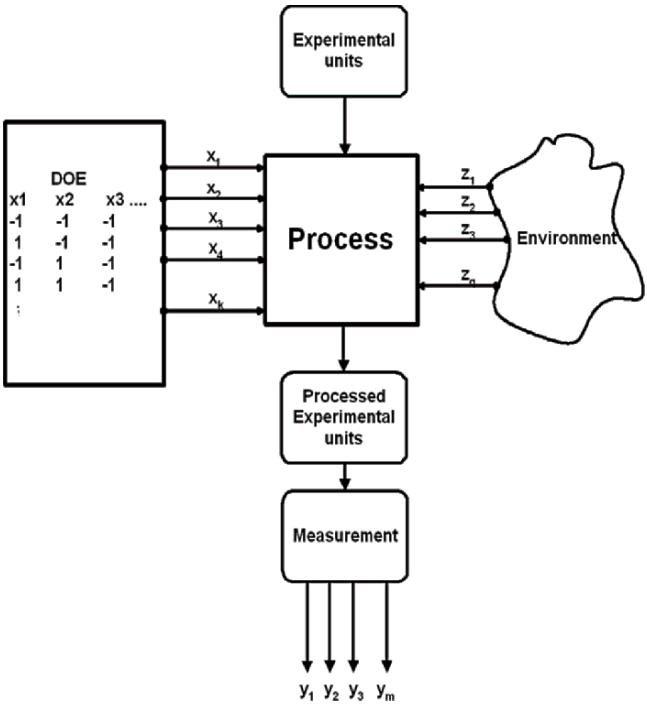


Figure 1.3. A process seen as a black box

noisy function by means of simpler approximating functions that are valid over a small region using designed experiments. By moving the operating conditions of a process using a sequence of experimental designs, process improvement is achieved.

As all important ideas in science and engineering, RSM did not emerged from a vacuum. Least squares estimation and regression analysis existed thanks to Gauss and, later, Galton, and it is possible they were applied in production or industrial design at some point of the XIX century, although there is no such historical evidence as far as we know. The idea of studying an engineering system based on approximating functions dates back to the end of the XIX century, when the Russian mathematician and engineer Chebyshev worked in finding an approximating function for a given function or trajectory that a mechanical artifact should follow. His work continued a long tradition in the area of function approximation and Fourier analysis. This lead to the concept of best

uniform approximation and his celebrated Chebyshev polynomials. The function to approximate is assumed known, so no statistical issues arise [2].

Industrial experimentation for the purposes of product design started in a more systematic way at the end of the XIX century, with Edison's company – perhaps the first industrial research and development center – and, almost in parallel to it, the work of Frederick Taylor, the "father" of industrial engineering, on metals for cutting tools. It is at least curious for a modern Industrial Engineer familiarized with experimental design techniques to look at Taylor's work and see the tremendous amount of time and energy that could have been saved if DOE techniques would have been available to Taylor at the time. The same can be said about Edison and all the engineers and inventors that worked for him. For this reason, this era can be called the "heroic era" of experimental design in industry.

Closer in time to the Box-Wilson paper, ideas on process optimization from a statistical perspective were put forward by Hotelling [71], who discusses applying these ideas to the optimization of alloys. Hotelling considered a single factor problem (he only speculated about two and more factors), trying to derive the distribution and statistical properties of the optimum obtained through a polynomial response fitted via least squares. Interestingly, he seems to anticipate some of the ideas developed later by Box and Wilson when he mentions the importance of conducting experiments in stages, where "A large-scale investigation should be preceded by a smaller one designed primarily to obtain information for use in designing the large one. The small preliminary investigation may well in turn be preceded by a still smaller pre-preliminary investigation, and so on" [71]. At about the same time, the important work of Fisher on the analysis and design of factorial experiments in agriculture in the first half of the XX century [55] led the way to the application and adaptation of these techniques in industry for the purpose of optimizing production processes. It was from this basis that the paper by Box and Wilson took off. Their work and the subsequent work in RSM that originated from it has had the greatest impact in industrial practice.

At about the same time than the Box and Wilson paper appeared, a parallel development took place in the Mathematical Statistics field. The paper by Robbins and Monro [135] on Stochastic Approximation of the root of a function has spurred an impressive body of literature ever since. A year after the

Robbins-Monro paper appeared, Kiefer and Wolfowitz[5] [79] showed a convergence proof for a search technique for the maximum of a regression function based on stochastic-approximation, a proof that is nonexistent in the case of RSM due to its much more general assumptions. Later authors extended these results to several dimensions (Keifer and Wolfowitz dealt with the case of a single dimension), and recent work has shown the many connections between "stochastic optimization" and other sequential estimation techniques used in Adaptive Control (for a nice recent discussion, see Spall [145]). A recent technique based on this school of thought is Spall's Simultaneous Perturbation Stochastic Approximation (SPSA) methods, which have been reported to be useful in the optimization of noisy "black box" systems.

Optimization techniques based on stochastic approximation ideas have a radically different point of view than RSM techniques. Instead of a sequence of models obtained from carefully designed experiments over a small experimental region that is changed in view of the experimental results, stochastic approximation methods concentrate on a more mathematically rigorous algorithmic search approach to process optimization, trying to find good convergence properties under the fewest possible assumptions. The minimum set of assumptions needed for proving convergence is usually too strict to be satisfied in real-life industrial experiments. However, this does not imply that these methods will never work in practice. For example, recent computational experience with SPSA methods indicates that for certain applications, like simulation optimization, they should certainly be considered as an important optimization tool. SPSA and other related techniques will be discussed in Chapter 13.

In the earlier 1980's, G. Taguchi came to the USA (he did work at AT&T Bell Labs) and popularized his ideas about noise factors and robustness. Initially, his ideas were accepted in an uncritical manner by some of his followers. In the late 80's and early 90's, Taguchi's ideas were discussed and criticized in the major Engineering Statistics journals of the US, and this resulted in a consensus, among the leading experts, that Taguchi's ideas were innovative (in particular, the concept of a noise factor) but some of the statistical methods he suggested to reach the ultimate goals of Robust Parameter Design were inappropriate. Since then, more sound statistical methods for Robust Parameter

[5]Kiefer and Wolfowitz later developed in great detail the mathematical theory of optimal experimental design, which we cover in Chapter 5.

Design have been proposed in the literature. We will look at some of them in Part IV.

1.3 Some Basic Ideas in Response Surface Methods

We now turn in more detail to describe the mainstream process optimization ideas in the Statistics and Quality Engineering literature, namely, RSM. In order to be more precise about the general aims of RSM, suppose we have a single response which is a function of k controllable factors. After each experiment, we observe:

$$y = \eta + \varepsilon$$

where

$$\eta = f(x_1, x_2, \ldots, x_k)$$

is the true response value, unobservable directly, since we hypothesize we always observe η in conjunction with additive noise or error ε. The function f is unknown and directly unaccessible to us, and one goal in design of experiments is to model or **characterize** the function f.

The error or noise term ε is usually assumed distributed with mean zero and constant variance σ_ε^2, which implies that

$$E[y] = \eta$$

is the mean response value and $y - \eta = \varepsilon$ is the error term, a random variable. The error term models, in a collective manner, all other potential effects that we do not understand or know (we will not consider noise factors until Part IV). This is a classical form of a statistical model: it is composed by a part we can explain and a part we cannot explain, which we would like to make as small as possible in some well-defined way.

We will assume, for the most part in this book, that all the controllable factors are quantitative. Let x_i denote the **level** or value of factor i, $i = 1, \ldots, k$. This is the setting we, as experimenters, decide for this factor in a particular experimental run. Experimental design techniques suggest that these levels be scaled or **coded** into a unitless value, typically on the -1 to $+1$ range, in such a way that an experimenter can pick an experimental design from a "catalog" of designs, being able to compare alternative designs (other advantages of coding will be discussed later). The most usual coding convention is the orthogonal

convention given by:

$$x_i = \frac{\xi_i - M_i}{R_i/2}, \qquad i = 1, 2, \ldots, k \tag{1.1}$$

where ξ_i is the level or value of controllable factor i in original units of measurement, M_i is the midpoint of the range of values for factor i in original units, and R_i is the range of values over which we will vary factor i in original units. In vector notation, a coding transformation will have the general simple form

$$\boldsymbol{x} = \boldsymbol{S}^{-1}(\boldsymbol{\xi} - \boldsymbol{M})$$

where $\boldsymbol{\xi}$ is a $k \times 1$ vector of controllable factors, $\boldsymbol{M}$ is a $k \times 1$ vector that determines a change in location and $\boldsymbol{S}$ is a $k \times k$ matrix which determines a change in the scale of the factors. For the particular case of coding (1.1) we have that

$$\boldsymbol{S} = \frac{1}{2}\begin{pmatrix} R_1 & 0 & 0 & \cdots & 0 \\ 0 & R_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & R_k \end{pmatrix}$$

and $\boldsymbol{M}$ contains the midpoints of the factors in original units. This coding convention maps the highest value of each controllable factor into the $+1$ level, the lowest value is mapped into the -1 level, and the midpoint of the range is mapped into the 0 (zero) level. For this reason, coding (1.1) will be referred to as the "$(-1,1)$" coding convention.

Evidently, the transformation (1.1) is not unique, and others have been proposed and will be discussed at certain points in this book. However, this is the most frequent coding convention in practice. Coding conventions of this type change the metric of the experimental design and with this they change the properties of the parameter estimates associated with the corresponding model. These ideas will be more fully explained in Chapter 3.

An experimenter is usually not interested in the complete k-dimensional space defined by the factors ξ (in original units). This complete space where experiments can in principle be conducted is referred to as the **region of operation** or region of operability. In some chemical or manufacturing processes, for example, a combination of controllable factors may be unsafe for the operation of the process or may be well-known to produce scrap. This imposes a constraint in the factor space which in turn poses special problems for experimental

design. We discuss this issue in Chapter 5. In addition to the region of operation we have the **region of interest**, that the experimenter actually wishes to explore. In industrial experiments, it is within this region where we should request good properties of the fitted response model, with the understanding that extrapolating outside this region is not an immediate goal. In this case, the literature mainly highlights the two most common regions of interest: a "cuboidal" region and a "spherical" region, assuming these regions do not touch the boundaries of the region of operation. The region of interest may still be a proper subset of the region of operation but may be close to a boundary of the latter, which imposes a constraint in the region of interest itself.

In the area of **computer experiments** (see Chapter 14), the region of interest is frequently equal to the region of operation, and this calls for different DOE and model building strategies, as we wish to have adequate response model properties throughout the region of operation. In either case, industrial or computer experiments, we point out that the region of operation may not necessarily have a cuboidal form, since in most real-life experiments there might be constraints on the operation of a process.

We should further distinguish the region of interest from the actual **experimental region**. The experimental region is usually defined by the ranges R_i of the controllable factors once the experiment has been designed. There is the intuitive notion that the experimental region should resemble the region of interest. Thus, for example, if interest is in a cuboidal region, then a cuboidal experimental design should be used. This is easier to do for cuboidal regions than for spherical regions of interest, for which only some designs are truly spherical in the way they distribute the design points on the factor space.

The center or origin of the experimental region corresponds to the point whose coordinates are equal to zero in coded units. These can represent "baseline" operating conditions. In industrial experiments, the experimental region in RSM is quite small compared to the region of operation, has a geometry that resembles the region of interest, and typically moves as we explore new regions within the region of operation. The center "point" in coded units refers to the center of the *current* experimental region.

The main goal of RSM techniques is **process optimization**. This could imply, for example, to minimize the cost of operation of a production process, to minimize the variability of a quality characteristic, to maximize the yield in a

chemical process, or to achieve desired specifications for a response. Evidently, multiple responses of interest are usually considered in practical problems.

In order to optimize an industrial process, RSM methods suggest to build a parametric model for the expected response using designed experiments. These models should be **local** approximations, valid in a small experimental region. For example, a first order polynomial approximation in two ($= k$) controllable factors is given by:

$$E[y] = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{1.2}$$

A second order polynomial approximation, useful in a region where the response is "curved" (we will say that "**curvature** is present" in such a case) is given instead by:

$$E[y] = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2. \tag{1.3}$$

Here, the β's are unknown parameters that must be estimated from experimental data. The β_i parameters are estimating the main effects of each factor, that is, the average change in response as we move factor i by a unit[6]. The parameter β_{12} estimates a **two factor interaction**, that is, the additional effect that the response will experience on top of the main effects for the two factors. The effect of adding a 2-factor interaction to a first order model is to curve the function in the form of a hyperbola. Finally, we have the β_{jj} parameters, which are called the "**pure quadratic**" effects, to distinguish them from 2-factor interactions, which are also quadratic but due to the mixed effect of two factors. If added to a linear model they will induce parabolic curvature on the function.

More generally, a **first order polynomial model** in k factors is

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon$$

which can be written as

$$y(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} + \varepsilon$$

[6]In DOE, main effects are really defined as the average change as the factor is moved over a two unit range, from -1 to 1.

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are $k \times 1$ vectors. Similarly, a **second order polynomial model** for k factors is

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \sum_{i=1}^{j-1} \sum_{j=2}^{k} \beta_{ij} x_i x_j + \varepsilon$$

which can be written

$$y(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x} + \mathbf{x}'\boldsymbol{\Xi}\mathbf{x} + \varepsilon \tag{1.4}$$

where $\boldsymbol{\Xi}$ is a symmetric $k \times k$ matrix that contains the β_{ii} parameters in the diagonals and $1/2\beta_{ij}$ in the (i, j) off-diagonal positions. Notice how the polynomial models we are proposing are always linear *in the parameters*. In contrast, the second order model is non-linear in the controllable factors.

The justification of a polynomial as an approximating function comes from a simple Taylor series expansion of the unknown function f (assumed twice differentiable) around the center of the current experimental region. Suppose $k = 1$. Then

$$y = f(x_1) + \varepsilon,$$

and we have that

$$f(x_1) = f(0) + x_1 \frac{df}{dx_1}\bigg|_{x_1=0} + \frac{1}{2} x_1^2 \frac{d^2 f}{dx_1}\bigg|_{x_1=0} + \cdots + \frac{x^n}{n!} \frac{d^n f}{dx_1}\bigg|_{x_1=0}$$

Neglecting terms of third and higher order we get:

$$f(x_1) \approx f(0) + x_1 f'(0) + x_1^2 f''(0)/2$$

from where we can see that if we fit a second order polynomial in one factor, β_0 corresponds to the intercept of the function at the center ($f(0)$), β_1 corresponds to the slope of the function ($f'(0)$) at the design center, and β_{11} corresponds to a scaled measure of the curvature of the function at the center (as given by the Hessian, $f''(0)$). The parameters are usually fit using ordinary least squares.

More generally, for a function of k factors, $f(\mathbf{x})$, assumed twice differentiable, we have that

$$f(\mathbf{x}) \approx f(\mathbf{0}) + \nabla f(\mathbf{0})'\mathbf{x} + \frac{1}{2}\mathbf{x}'\nabla^2 f(\mathbf{0})\mathbf{x}. \tag{1.5}$$

where $\nabla f(\mathbf{0})$ is the gradient of f evaluated at the origin and $\nabla^2 f(\mathbf{0})$ is the Hessian of f evaluated at the origin. Comparing (1.4) and (1.5) we see that by

fitting (1.4), we are estimating the gradient and the Hessian of the function of interest at the origin of the current experimental region. It should be pointed out the differentiability assumption on which this Taylor series argument rests. It implies that the functions we are expecting to find, although unknown, will be sufficiently "smooth".

A key insight in the Box and Wilson paper [29] was to suggest that for a poorly understood process, a model of the type (1.2) will typically be appropriate, since the process will likely be operating far from its optimum, in a region where "curvature" should not exist. A model such as (1.3) will typically prove to be useful in a region in the space of the controllable factors closer to an optimum, since it allows to model such curvature in the response (see Figure 1.4). These and later authors proposed a means to move the process from a region of little curvature to one where there is curvature, suggested specific types of designed experiments to be used at each stage and specific tests to decide between model (1.2) and model (1.3). Although there was no specific concern in the RSM literature for the application or use of the ideas of non-linear programming (see Appendix C) in noisy processes, the main reasoning of achieving a local optimum through a search process mimics the underlying goal of deterministic optimization techniques.

1.3.1 Steps in a Classical RSM Study

From its inception, the application of RSM techniques was seen as usually divided in 3 stages of phases:

1 **Factor screening**.- find the active or important factors among a wider set of possible controllable factors;

2 **steepest ascent/descent searches**.- used to "move" the operating conditions of a process to a region of experimentation (within the region of operation) where there is a better performance;

3 **canonical analysis**.- used to model curvilinear relationships and find stationary points of the response functions.

The Factor screening phase is preliminary to an actual optimization study, and is based on the Pareto-like idea that a small number of controllable factors will have the most important effect on a response of interest. This idea sometimes goes by the name **Sparsity of effects principle**, and its practical value in

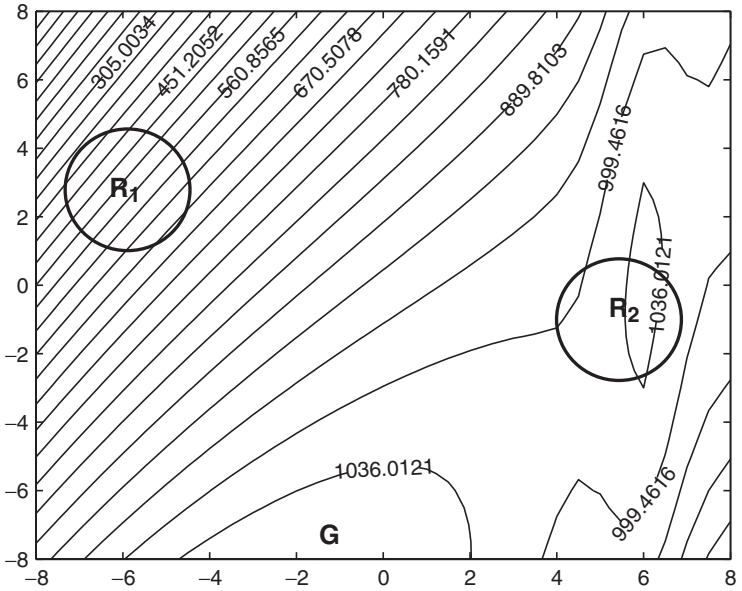


Figure 1.4. Contours of a hypothetical response as a function of two controllable factors. RSM attempts to move the process from a region R_1 of little curvature (where model (1.2) is appropriate) to a region R_2 where model (1.3) would provide better fit. Notice how R_2 is near a local maximum, the global maximum in the region plotted being located near point G

the simplification of an experiment cannot be exaggerated. This and the other two stages will be treated in greater detail in Part II of this book.

The steepest ascent (or descent) procedure consists in changing the operating conditions of the process in the direction of the gradient of the response surface. We discuss the classic use of steepest ascent in RSM in detail in Chapter 2, and at a more advanced level in Chapter 6. This stage is conducted until significant curvature in the response is detected. Once curvature is detected, a second order model is fit as a local approximation of the response. Then, canonical analysis of the quadratic response surface model consists in finding the stationary point of the response and its characteristics. Classic canonical analysis in RSM is fully described in Chapter 4. We introduce it here with the following example that serves also to illustrate a response surface experiment conducted in a traditional manufacturing process. The example assumes basic knowledge of linear regression techniques (see Appendix A).

Example. Optimization of a machining process. Consider an experiment conducted to determine the life of a cutting tool. The effects of three variables,

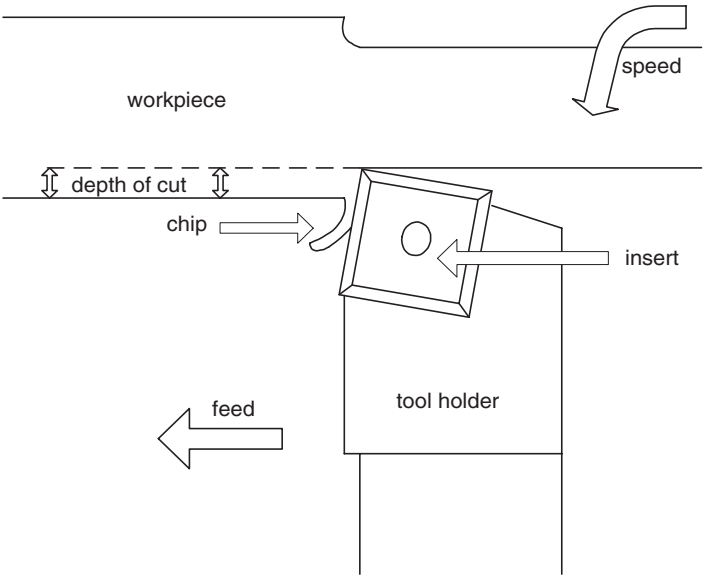


Figure 1.5. A typical machine tool showing the three main process variables: speed of cut, depth of cut, and feed

cutting speed, feed rate, and depth of cut on the tool life performance are of interest[7]. Figure 1.5 illustrates the three variables in a typical lathe machine. The speed was measured in units of surface feet per minute (sfm's), feed in inches per revolution (ipr), depth of cut in inches, and tool life in minutes. Tool life was determined by the first occurrence of 0.015 inch uniform flank wear, 0.004 inch crater depth, 0.030 inch localized wear, or catastrophic failure. The three controllable factors were scaled according to the coding convention (1.1): $x_1 = (speed - 725)/75$, $x_2 = (feed - 0.018)/0.008$, and $x_3 = (depth - 0.125)/0.075$. A central composite design, an experimental design we will discuss in detail in Chapter 5 was used. The experimental data is shown in Table 1.1. The experiment was run in a completely randomized way.

The following second order model was fit using least squares in the coded factors:

$$\widehat{y} = 6.56 - 5.99x_1 - 12.66x_2 - 4.51x_3 - 0.46x_1^2 + 8.68x_2^2 + 1.73x_3^2 + 4.8x_1x_2 + 1.82x_1x_3 + 1.80x_2x_3$$

[7]I thank Mr. Mike Delozier for providing the information related to this experiment [44].

Table 1.1. Experimental design and data for the machining example

Speed	**Feed**	**Depth of cut**	**Tool life**
650	0.010	0.050	49.4
800	0.010	0.050	22.0
650	0.026	0.050	10.0
800	0.026	0.050	3.2
650	0.010	0.200	32.2
800	0.010	0.200	13.5
650	0.026	0.200	1.4
800	0.026	0.200	0.5
650	0.018	0.125	9.1
800	0.018	0.125	3.0
725	0.010	0.125	27.5
725	0.026	0.125	2.9
725	0.018	0.050	12.3
725	0.018	0.200	4.2
725	0.018	0.125	7.1
725	0.018	0.125	8.6
725	0.018	0.125	4.2

with an estimated variance of $\widehat{\sigma}^2 = 2.187^2$ and a $R^2 = 0.988$. The residual plots (see Figure 1.6) look satisfactory. The fitted coefficients of the x_1^2 and x_3^2 terms are not significant; however, they were kept since the adjusted R^2 statistic improves very little (from 0.973 to 0.974) if these terms are deleted from the model, and the residuals look better behaved with these terms in the model. The lack of fit test (see Appendix A) showed that the fitted model exhibited no lack of fit. While transformations are usually necessary in life data, this was not the case in the present data set[8]. Writing the fitted model as in equation (1.4), we have that $\widehat{\boldsymbol{\beta}}' = (-5.99, -12.66, -4.51)$ and

$$\widehat{\boldsymbol{\Xi}} = \begin{bmatrix} -0.46 & 2.4 & 0.91 \\ 2.4 & 8.68 & 0.90 \\ 0.91 & 0.90 & 1.73 \end{bmatrix}.$$

[8]In fact, F. W. Taylor obtained a first order model relation for the logarithm of tool life as a function of speed in his experiments on high-speed steel tool materials in the early 1900s.

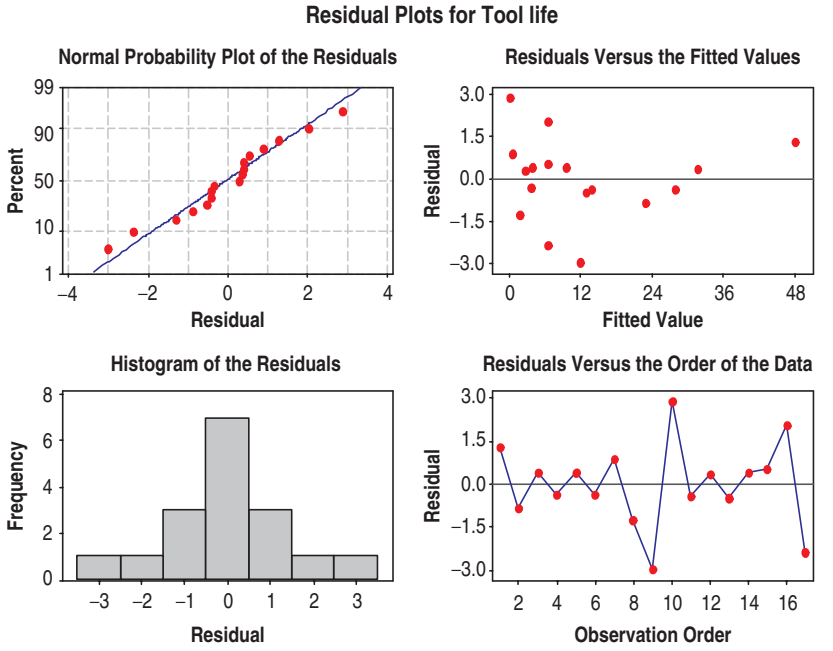


Figure 1.6. Residual plots for the fitted model, machining example

This eigenvalues of this matrix[9] are -1.2098, 1.732, and 9.4276, thus the fitted response is a saddle function which has a stationary point at $(-0.43, 0.72, 1.15)$ which is neither a maximum nor a minimum, and therefore, not very interesting in itself. The fitted function is displayed in Figure 1.7.

The previous analysis corresponds to "canonical analysis", which is essentially equivalent to unconstrained optimization of a second order model and determination of the nature of the fitted response through analysis of the fitted Ξ matrix and, if feasible, contour plots of the fitted response. Since the stationary point found is not a maximum (the goal was to maximize tool life), and the response has a good fit, *constrained* optimization could be conducted (see Chapter 4). ■

[9]See Appendix D.

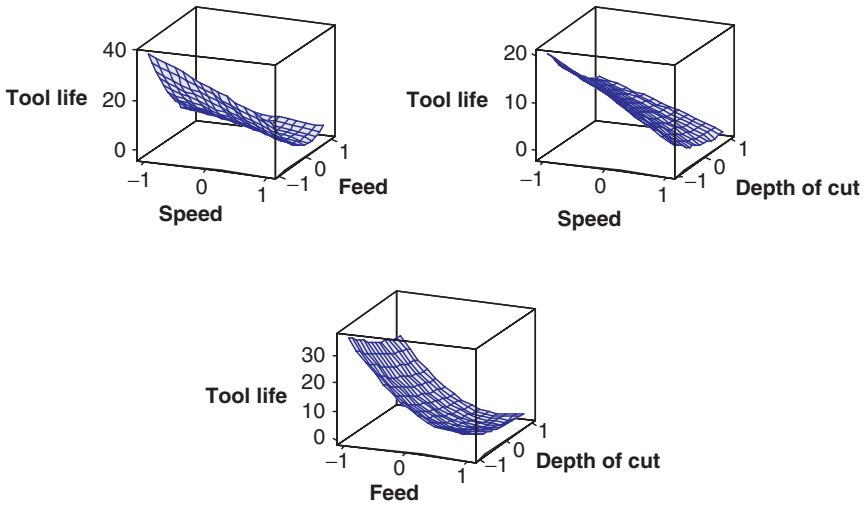


Figure 1.7. Fitted tool life response function (factors not shown held at their middle values of 0.0 in coded units)

1.4 Statistical Inference in Process Optimization: Frequentist and Bayesian Methods

It was thought by early RSM authors that it does not make sense to apply sophisticated optimization techniques to models fitted from data subject to large errors. Work conducted in the last 10 years has shown this statement not to be completely correct. If it were true, what is essentially an unconstrained optimization technique such as "canonical analysis" (see Chapter 4) should not be utilized. The statement is only true in the sense that it is important to consider the uncertainty in the data (and in general, all the uncertainties present) in the optimization problem. **Statistical inference** in RSM is very important in practice, and is one of the main themes in this book.

Experimental optimization of response surface models differs from classical optimization techniques in at least three ways:

1 Experimental optimization is an iterative process, that is, experiments conducted in one set of experiments result in fitted models that indicate where to search for improved operating conditions in the next set of experiments. Thus the coefficients of the fitted equations (or the form of the fitted

equations) may change during the optimization process. This is in contrast to classical optimization where the functions to optimize are supposed to be fixed and given.

2 The response models are fitted from experimental data that usually contains random variability due to uncontrollable or unknown causes. This implies that an experiment, if repeated, will result in a different fitted response surface model that might lead to different optimal operating conditions. Therefore, *sampling variability* should be considered in experimental optimization. In contrast, in classical optimization techniques the functions are deterministic and given.

3 The fitted responses are local approximations, implying that the optimization process requires the input of the experimenter (a person familiar with the process). This is in contrast with classical optimization which is always automated in the form of some computer algorithm.

To illustrate the kind of Statistical Inference problems that occur in process optimization, suppose we ran two separate DOEs, D_1 and D_2, on a process, and fit models m_1 and m_2 from each experiment. The optimum point $\boldsymbol{x}^*$ that achieves the optimal predicted response $\widehat{Y}(\boldsymbol{x}^*)$ will differ if obtained from m_1 than from m_2. Thus, there is sampling variability in $\boldsymbol{x}^*$ since it clearly varies as the data varies. The value of $\boldsymbol{x}^*$ obtained from a single experiment is simply a *point estimate* on the optimal operating conditions. A practical an important question is: how much will $\boldsymbol{x}^*$ vary due to sampling variability? Chapter 7 discusses frequentist methods of statistical inference in process optimization problems that address this kind of problems, both for unconstrained and for constrained optimization methods.

A different question, often asked by engineers is: what *probability* do we have that the response at $\boldsymbol{x}^*$, $y(\boldsymbol{x}^*)$ will satisfy some given tolerances? And how can we find $\boldsymbol{x}^*$ so that "yield" is maximized? In Part V we present **Bayesian optimization** approaches that address these and related questions. The Bayesian approach to process optimization allows to consider not only the uncertainty in the data but also uncertainties that exist due to: 1) not knowing the model parameters of a given model form; 2) not knowing the true *form* of the model; and 3) variability created by noise factors in the model.

1.5 Some Basic Ideas in Experimental Design

Before ending this introductory chapter, it is worth emphasizing the value of factorial experiments compared to a strategy where only one factor is varied at a time[10]. There are also some misconceptions about the latter strategy that should be pointed out.

To illustrate the use of a one-factor at a time (OFAT) experimental strategy, consider the left display on Figure 1.8. The figure shows a hypothetical quadratic function which we wish to maximize (the contours give increasingly higher function values as we move inside the parabolas) by varying two continuous factors. Let us suppose we start varying x_2, using increasingly larger values (or levels) of this factor, while keeping x_1 fixed at a value of zero. In this case we will move the process "up" a hill which will allow us to find a local optimum in that direction. The local optimum seems to be close to $x_2 = 0$. Thus, we now repeat the procedure, varying x_1 only while keeping x_2 at its (local) optimal value of zero. A similar situation will occur: we will find another "hill" in this direction, concluding that a value close to $x_1 = 0$ is the (local) optimum in this direction. Thus, we come to the (erroneous) conclusion to run the process at the settings (0, 0), which evidently are very far away from the region where the global maximum is located.

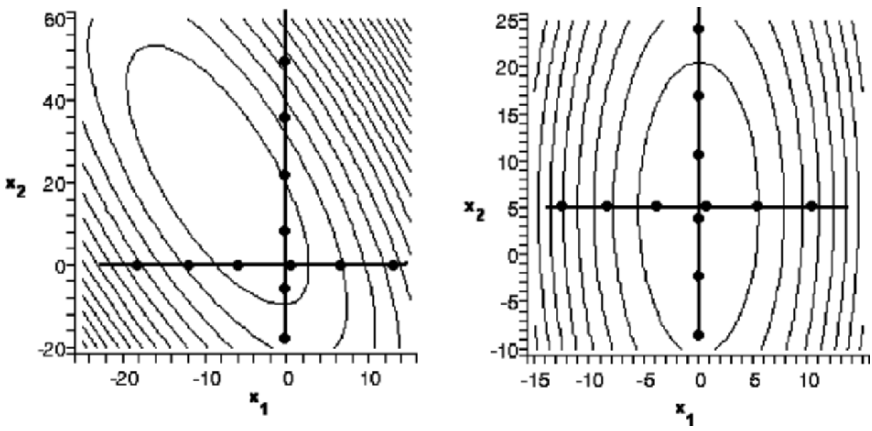


Figure 1.8. A case when a one factor at a time strategy does not work well (left) and a case when it works well (right)

[10]Readers familiar with the basic notions of DOE can skip this section without loss of continuity.

The function on the left panel in Figure 1.8 has a significant $\beta_{12}x_1x_2$ term, i.e., a significant two factor interaction. By varying only one factor at a time (in the way shown in Figure 1.8), we "miss" this interaction, which has the effect of rotating the paraboloid (note how as the interaction gets larger, the paraboloid form will eventually be lost, and the function will turn instead into a saddle function). In contrast, the right panel on Figure 1.8 shows a situation in which an OFAT strategy *will* work well, finding the global optimum rapidly. The function plotted on this figure does not have an interaction at all, therefore, the main axis of the paraboloid coincides with the coordinate axis of the factors. Which case is more common in practice? In practice, the probability of having to deal with a function that is "nicely oriented" in such a way that there are no interactions is very small, with the probability going further down as the number of factors k increases. Thus, in practice, since the function to optimize is unknown, it is much safer to assume there might be interactions and experiment accordingly. We note, in passing, that if the function can be represented by a second order polynomial and interactions are present, *rotating* the function to align the coordinate axis according to the main axis of the quadratic function, by means of an algebraic transformation, will simplify the resulting function and will facilitate its analysis as it will eliminate the interactions in the rotated coordinates. This was noted by Box and Wilson, and is the basis of their "canonical analysis", which we discuss in Chapter 4. The effects of rotations on the properties of a DOE have also been studied for a long time. We look at the *rotatability* properties of a DOE in Chapter 5.

An experimental strategy that is adequate if there are significant interactions is the **factorial design** strategy. In a factorial design, *many* factors are varied from run to run. In industrial practice, 2-level factorial experiments are frequently run. Figure 1.9 shows a 2-level factorial run in the case of the hypothetical function depicted on the left of Figure 1.8. Comparison of the four response values at each of the four test conditions will indicate the need to "move the process" in a direction close to that given by the arrow. The arrow points in the direction of **steepest ascent** from the center of the design. In Chapter 2 we discuss how to perform steepest ascent based on experimental data. We will discuss 2-level factorial experiments in Chapter 3. Experimental designs that allow to fit the second order model (1.4), such as the central composite design shown in the machining example, will be discussed in Chapter 5.

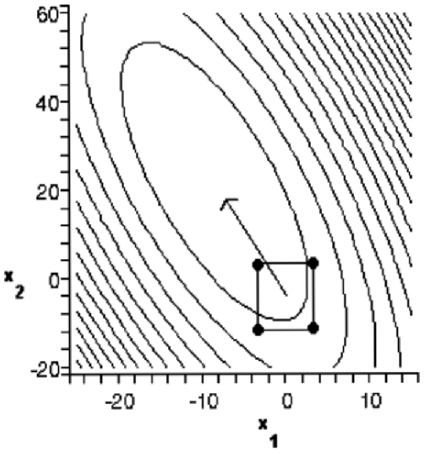


Figure 1.9. A 2-level factorial strategy showing the direction of steepest ascent

Shall we *never* use an OFAT strategy? The answer is certainly *never* use it in the way depicted in Figure 1.8. There has been some recent work where a special modification of OFAT designs, **Adaptive OFAT**, is shown to be useful and inexpensive (we will discuss this approach in Chapter 3). Adaptive OFAT methods attempt to "exploit" the interactions, and do not move simply as in Figure 1.8 but use a sequence of 2-level OFAT designs that allow to "backtrack" to the best point found thus far. Despite this interesting new development, we note that, as a general principle in experimental design, a factorial design should (almost always) be preferred. We discuss some further basic principles of experimental design, such as **randomization** and **blocking**, in Chapters 3 and 5.

PART II

ELEMENTS OF RESPONSE SURFACE METHODS

Chapter 2

OPTIMIZATION OF FIRST ORDER MODELS

One should not multiply explanations and causes unless it is strictly necessary
—"William of Bakersville" in Umberto Eco's *In the Name of the Rose*[1]

In Response Surface Methods, the optimal region to run a process is usually determined after a sequence of experiments is conducted and a series of empirical models are obtained. As mentioned in Chapter 1, in a new or poorly understood process it is likely that a first order model will fit well. The Box-Wilson methodology suggests the use of a steepest ascent technique coupled with lack of fit and curvature tests to move the process from a region of little curvature to one where curvature – and the presence of a stationary point – exists. In this chapter we discuss, at an elementary level, steepest ascent/descent methods for optimizing a process described by a first order model. We will assume readers are familiar with the linear regression material reviewed in Appendix A. More advanced techniques related to exploring a new region that incorporate the statistical inference issues into the optimization methods are discussed in Chapter 6.

[1]Paraphrasing a non-fictional Franciscan monk of the epoch, William of Ockham (1285–1347): *Essentia non sunt multiplicanda praeter necessitatem* (Entities should not be multiplied beyond necessity), a statement that became to be known as "Ockham's razor".

2.1 New Region Exploration

A first order model will serve as a good local approximation in a small region close to the initial operating conditions and far from where the process exhibits curvature. A fitted first order polynomial model in k factors has the form

$$\widehat{y} = \widehat{\beta}_0 + \sum_{i=1}^{k} \widehat{\beta}_j x_j.$$

Experimental designs used for fitting this type of models are discussed in more detail in Chapter 3. Usually, a two-level factorial experiment is conducted with repeated runs at the current operating conditions which serve as the origin of coordinates in coded factors.

The idea behind this region exploration "Phase" of RSM is to keep experimenting along the direction of steepest ascent (or descent, as required) until there is no further improvement in the response. At that point, a new fractional factorial experiment with center runs is conducted to determine a new search direction. This process is repeated until at some point significant curvature in $\widehat{y}$ or lack of fit in the 1st order model will be detected. In the new region, the operating conditions $x_1, x_2, \ldots, x_k$ are close to where a stationary point (and hopefully, a local maximum or minimum, as required) of y occurs. When significant curvature or lack of linear fit is detected, the experimenter should proceed to fit and optimize a higher order model, in what Box and Wilson called "canonical analysis". Figure 2.1 illustrates a sequence of line searches when seeking a region where curvature exists in a problem with 2 factors (i.e., $k = 2$). In practical problems, more than 3 iterations of steepest ascent/descent are rare.

There are two main decisions we must make in the region exploration phase of RSM:

1 determine the search direction;

2 determine the length of the step to move from the current operating conditions.

These two decisions are linked, because a search in a direction where there is strong evidence the process will improve may proceed at larger steps than a search in a less reliable direction. In Chapter 6 we present some approaches on how to deal with these two aspects jointly.

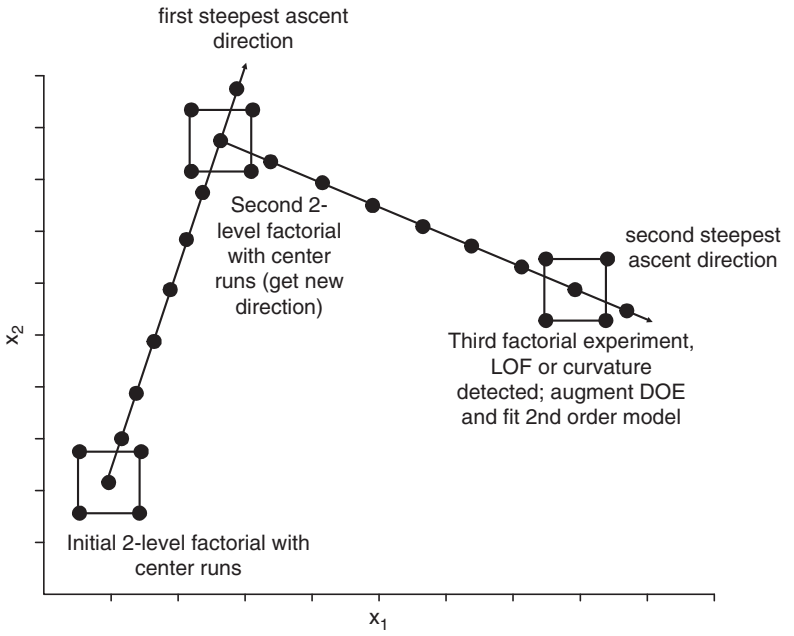


Figure 2.1. A sequence of line searches for a 2 factor optimization problem

2.2 Steepest Ascent/Descent Procedure

Suppose a first order model has been fitted and the model does not exhibit lack of fit due to second order terms. Then the direction of *maximum improvement* is given by:

1 $\boldsymbol{\nabla}\widehat{y}$, if the objective is to *maximize* y (this gives the steepest ascent direction);

2 $-\boldsymbol{\nabla}\widehat{y}$, if the objective is to *minimize* y (this gives the steepest descent direction).

The direction of the gradient is given by the values of the parameter estimates (excluding the intercept), that is, $\widehat{\boldsymbol{\beta}}' = \boldsymbol{b}' = (b_1, b_2, \ldots, b_k)$. Since the parameter estimates depend on the scaling convention for the factors, the steepest ascent (descent) direction is also scale-dependent. That is, two experimenters using different scaling conventions will follow different paths

for process improvement. We note, however, that the region of the search, as given by the signs of the parameter estimates, does not change with scale[2].

Let us consider without loss of generality a maximization problem. The coordinates of the factor settings on the direction of steepest ascent separated a distance ρ from the origin are obtained from solving:

$$\begin{aligned} \max \quad & b_1x_1 + b_2x_2 + \cdots + b_kx_k = \boldsymbol{b}'\boldsymbol{x} \\ \text{subject to:} \quad & \sum_{i=1}^{k} x_i^2 \leq \rho^2 \quad \text{or} \quad \boldsymbol{x}'\boldsymbol{x} \leq \rho^2. \end{aligned}$$

Unconstrained optimization of the first order model will evidently provide an unbounded solution, hence the constraint is added. The value of ρ is the step size, which is user-defined. A value of $\rho = 1$ will give a point close to the edge of the experimental region where the model was fit (in coded factors). To solve the maximization problem, form the Lagrangian (see Appendix C for general results on optimality conditions):

$$\text{maximize} \quad L = \boldsymbol{b}'\boldsymbol{x} - \lambda(\boldsymbol{x}'\boldsymbol{x} - \rho^2)$$

where λ is a Lagrange multiplier. Compute the partials and equate them to zero

$$\frac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{b} - 2\lambda\boldsymbol{x} = \boldsymbol{0}$$

$$\frac{\partial L}{\partial \lambda} = -(\boldsymbol{x}'\boldsymbol{x} - \rho^2) = 0.$$

These two equations have two unknowns (the vector $\boldsymbol{x}$ and the scalar λ) and thus can be solved yielding the desired solution:

$$\boldsymbol{x}^* = \rho\frac{\boldsymbol{b}}{||\boldsymbol{b}||}$$

or, in non-vector notation:

$$x_i^* = \rho\frac{b_i}{\sqrt{\sum_{j=1}^{k} b_j^2}} \quad i = 1, 2, \ldots, k. \tag{2.1}$$

From this equation, we can see that any multiple ρ of the direction of the gradient (given by the unit vector $\boldsymbol{b}/||\boldsymbol{b}||$) will lead to points on the steepest

[2]For a recent proposal on a scale-independent direction finding method, see Section 6.3.

ascent direction. For steepest descent, we use instead $-b_i$ in the numerator of the equation above.

An equivalent and simple approach to finding the points on the steepest ascent path is as follows [110]:

1 Choose a step size in one of the controllable factors, say Δx_j. This may be the factor we are most comfortable varying or simply the one with the largest $|b_j|$.

2 Take a proportional step in all other factors, i.e.,

$$\frac{\Delta x_i}{b_i} = \frac{\Delta x_j}{b_j}, \quad i = 1, 2, \ldots, k;\ i \neq j$$

from which

$$\Delta x_i = \frac{b_i}{b_j}\Delta x_j, \qquad i = 1, 2, \ldots, k;\ i \neq j. \tag{2.2}$$

These two approaches are equivalent since the second approach will result in a point on the path of steepest ascent located at a distance $\rho = \sqrt{\sum_{i=1}^{k} \Delta x_i^2}$ (coded units) from the origin.

Starting at some given operating conditions (the origin in coded units), the recommended practice is to run a 2-level factorial experiment with center points replicated to allow to test for lack of fit and for curvature (see Appendix A). If there is evidence of lack of fit (LOF) or of curvature, we should add points to the 2-level factorial so that a second order model is estimable. With the second order model, we can then estimate the location of a local optima more precisely, as explained later in Chapter 4. We now illustrate the steepest ascent procedure.

Example. Optimization of a Chemical Process. It has been concluded (perhaps after a factor screening experiment) that the yield (y, in %) of a chemical process is mainly affected by the temperature (ξ_1, in °C) and by the reaction time (ξ_2, in minutes). Due to safety reasons, the region of operation is limited to

$$50 \leq \xi_1 \leq 250$$
$$150 \leq \xi_2 \leq 500$$

The process is currently run at a temperature of 200 °C and a reaction time of 200 minutes. A process engineer decides to run a 2^2 full factorial experiment

with factor levels at

$factor$	low	$center$	$high$
X_1	170	200	230
X_2	150	200	250

Five repeated runs at the center were conducted to assess lack of fit (see Appendix A for details on lack of fit tests). The orthogonally coded factors are

$$x_1 = \frac{\xi_1 - 200}{30} \qquad \text{and} \qquad x_2 = \frac{\xi_2 - 200}{50}.$$

The experimental results are shown in Table 2.1. The corresponding ANOVA table for a first order polynomial model is given in Table 2.2.

Neither the single degree of freedom test of curvature nor the lack of fit test indicate a problem with the model. Furthermore, there is evidence that the first

Table 2.1. First 2-level experimental design, chemical experiment example

x_1	x_2	ξ_1	ξ_2	Y(=yield)
−1	−1	170	150	32.79
1	−1	230	150	24.07
−1	1	170	250	48.94
1	1	230	250	52.49
0	0	200	200	38.89
0	0	200	200	48.29
0	0	200	200	29.68
0	0	200	200	46.50
0	0	200	200	44.15

Table 2.2. ANOVA table for first DOE, chemical example

Source	SS	d.o.f.	MS	F_0	P-value
Model	503.3035	2	251.6517	4.810	0.0684
Curvature	8.1536	1	8.1536	0.1558	0.7093
Residual (Error)	261.5935	5	52.3187		
Lack of Fit	37.6382	1	37.6382	0.6722	0.4583
Pure Error	223.9553	4	55.9888		
Total	773.0506	8			

order model is significant. The fitted equation in coded factors, using ordinary least squares (OLS), is

$$\widehat{Y} = 39.57 - 1.2925x_1 + 11.14x_2$$

Diagnostic checks (see Appendix A) show conformance to the regression assumptions, although the R^2 value (0.6580) is not very high.

To *maximize* $\widehat{y}$, we use the direction of steepest ascent. Suppose we select $\rho = 1$, since a point on the steepest ascent direction distanced one unit (in the coded units) from the origin is desired. Then from equation (2.1), the coordinates of the factor levels for the next run are given by:

$$x_1^* = \frac{\rho b_1}{\sqrt{\sum_{j=1}^2 b_j^2}} = \frac{(1)(-1.2925)}{\sqrt{(-1.2925)^2 + (11.14)^2}} = -0.1152$$

and

$$x_2^* = \frac{\rho b_2}{\sqrt{\sum_{j=1}^2 b_j^2}} = \frac{(1)(11.14)}{\sqrt{(-1.2925)^2 + (11.14)^2}} = 0.9933.$$

This means that to improve the process, for every $(-0.1152)(30) = -3.456\,^\circ\text{C}$ that temperature is varied (decreased), the reaction time should be varied by $(0.9933)(50) = 49.66$ minutes.

Alternatively, we could have used instead the procedure that lead to expression (2.2):

1 Suppose we select $\Delta\xi_2 = 50$ minutes. This can be based on process engineering considerations. It could have been felt that $\Delta\xi_2 = 50$ does not move the process too far away from the current region of experimentation. This makes sense since the R^2 value of 0.6580 for the fitted model is quite low, providing a steepest ascent direction not very reliable.

2 $\Delta x_2 = \frac{50}{50} = 1.0$.

3 $\Delta x_1 = \frac{-1.2925}{11.14} = -0.1160$.

4 $\Delta\xi_2 = (-.1160)(30) = -3.48\,^\circ\text{C}$.

Thus the step size in original units is $\Delta\boldsymbol{\xi}' = (-3.48\,^\circ\text{C}, 50 \text{ minutes})$. ■

To conduct experiments along the direction of maximum improvement, we just continue selecting operating conditions using the same step size as selected before, adding the step $\Delta\boldsymbol{\xi}$ to the last point on the steepest ascent/descent direction $\boldsymbol{\xi}_i$.

1 Given current operating conditions $\boldsymbol{\xi}_0' = (\xi_1, \xi_2, \ldots, \xi_k)$ and a step size $\Delta\boldsymbol{\xi}' = (\Delta\xi_1, \Delta\xi_2, \ldots, \Delta\xi_k)$, perform experiments at factor levels $\boldsymbol{\xi}_0 + \Delta\boldsymbol{\xi}, \boldsymbol{\xi}_0 + 2\Delta\boldsymbol{\xi}, \boldsymbol{\xi}_0 + 3\Delta\boldsymbol{\xi}, \ldots$ as long as improvement in the response y (decrease or increase, as desired) is observed.

2 Once a point has been reached where there is no further improvement, a new first order experiment (e.g. a 2-level factorial) should be performed with repeated center runs to assess lack of fit and curvature. If there is no significant evidence of lack of fit, the new first order model will provide a new search direction, and another iteration is performed. Otherwise, if either there is evidence of lack of fit or of (pure quadratic) curvature, the experimental design is augmented and a second order model should be fitted. That is, the experimenter should proceed with the next "Phase" of RSM.

Example. Experimenting along the direction of maximum improvement. Let us apply the steepest ascent procedure to the Chemical Experiment analyzed earlier. Recall that a first order model was significant, and did not show lack of fit or evidence of curvature. We proceed to move the process in the direction of steepest ascent, by using the step size computed earlier.

Step 1:

Given $\boldsymbol{\xi}_0 = (200\,^\circ\text{C}, 200 \text{ minutes})$ and $\Delta\boldsymbol{\xi} = (-3.48\,^\circ\text{C}, 50 \text{ minutes})$, perform experiments as shown on Table 2.3 (the step size in temperature was rounded to $-3.5\,^\circ\text{C}$ for practical reasons). Since the goal was to maximize y, the point of maximum observed response is $\xi_1 = 189.5\,^\circ\text{C}$, $\xi_2 = 350$ minutes. Notice that the search was stopped after 2 consecutive drops in response. This was done to reassure us that the mean response was *actually decreasing* and to avoid

Table 2.3. Illustration of a steepest ascent search

	ξ_1	ξ_2	x_1	x_2	y (=yield)
$\boldsymbol{\xi}_0$	200	200	0	0	
$\boldsymbol{\xi}_0 + \Delta\boldsymbol{\xi}$	196.5	250	−0.1160	1	56.2
$\boldsymbol{\xi}_0 + 2\Delta\boldsymbol{\xi}$	193.0	300	−0.2333	2	71.49
$\boldsymbol{\xi}_0 + 3\Delta\boldsymbol{\xi}$	189.5	350	−0.3500	3	75.63
$\boldsymbol{\xi}_0 + 4\Delta\boldsymbol{\xi}$	186.0	400	−0.4666	4	72.31
$\boldsymbol{\xi}_0 + 5\Delta\boldsymbol{\xi}$	182.5	450	−0.5826	5	72.10

stopping the search too early due to noise. This issue calls for a formal stopping rule for steepest ascent, a topic we discuss in detail in Chapter 6.

Step 2:

A new 2^2 factorial experiment is performed with $\boldsymbol{\xi}' = (189.5, 350)$ as the origin. Using the same scaling factors as before, the new scaled controllable factors are:

$$x_1 = \frac{\xi_1 - 189.5}{30} \qquad \text{and} \qquad x_2 = \frac{\xi_2 - 350}{50}$$

Five center runs (at $\xi_1 = 189.5, \xi_2 = 350$) were repeated to assess lack of fit and curvature. The experimental results are shown on Table 2.4. The corresponding ANOVA table for a linear model is shown on Table 2.5.

From the table, the linear effects model is significant and there is no evidence of lack of fit. However, there is a significant curvature effect, which implies

Table 2.4. Second 2-level factorial run, chemical experiment example

x_1	x_2	ξ_1	ξ_2	y (=yield)
-1	-1	159.5	300	64.33
1	-1	219.5	300	51.78
-1	1	158.5	400	77.30
1	1	219.5	400	45.37
0	0	189.5	350	62.08
0	0	189.5	350	79.36
0	0	189.5	350	75.29
0	0	189.5	350	73.81
0	0	189.5	350	69.45

Table 2.5. ANOVA table for the second DOE, chemical example

Source	SS	d.o.f.	MS	F_0	P-value
Model	505.300	2	252.650	4.731	0.0703
Curvature	336.309	1	336.309	6.297	0.0539
Residual	267.036	5	53.407		
Lack of fit	93.857	1	93.857	2.168	0.2149
Pure Error	173.179	4	43.295		
Total	1108.646	8			

that we have moved the process operating conditions to a region where there is curvature, so we should proceed to the next phase in RSM, namely, fitting and optimizing a second order model. ■

The discussants of the Box and Wilson paper [29] pointed out that no clear stopping criterion existed in the methodology, to which the authors agreed. Although the discussants were referring to the final phase of RSM, the same can be said about a steepest ascent search. Using a simple "first drop" stopping rule (i.e., stop if the response drops for the first time) may miss the true maximum in the steepest ascent direction, as the observed drop may be due to noise. Box and Wilson noted this, mentioning that the observed response values could be used to test for the difference of the mean response, and warned that the steepest ascent procedure is useful under low noise conditions. Formal tests of hypothesis that consider "noise" in the ascent procedure were developed by later authors, and we present them in Chapter 6. A Bayesian stopping rule, much in the sense of the discussion of the Box-Wilson paper, has been developed by Gilmour and Mead [61]. We discuss it in detail in Chapter 12.

Considerable work has been conducted since the Box and Wilson paper [29] on using search directions other than the steepest ascent direction, and on computing a confidence "cone" around a given direction. If the cone is too wide, this is evidence to proceed cautiously along such direction, so the cone provides some guidance about how large one should choose the step size along the steepest ascent direction other than selecting either ρ or Δx_j arbitrarily. Again, we refer to Chapter 6 for details. Directions other than steepest ascent, which is well-known in the non-linear programming literature to have "zig-zagging" problems, have been proposed as well. We review one such proposal in Chapter 6.

A final comment on the steepest ascent procedure as proposed in Box and Wilson refers to the assumed model. If a fractional factorial experiment at two levels is conducted, why neglecting two factor interactions? The argument in favor of this model is that in the initial stage of experimentation interactions will be dominated by main effects. This might be true for pure quadratic terms, but it is not clear why it should be true for two factor interactions. The argument reminds us a point in the controversy around the Taguchi methods (see Chapter 9). Taguchi considers *a priori* that some interactions are insignificant, and this has been criticized – fairly in our view – to be unjustified in general.

Likewise, in steepest ascent it seems that the first order model is assumed to fit the optimization technique (steepest ascent) because otherwise one would have to rely on relatively more complex nonlinear programming techniques. A two level factorial of enough resolution may be adequate to estimate two-factor interactions which can be utilized to optimize a process more efficiently, even at an earlier stage. The success of steepest ascent in practice, a evidenced by the many papers that have reported its application, is due to the fact that the technique will achieve some improvement – maybe considerable improvement – with respect to an earlier, non-optimized state of the process. That this improvement is as large as it could have been obtained or that it was found at the minimum cost (i.e., with a minimum number of experiments) has not been shown, and this point has not been emphasized much in the literature. Thus, there is room for improvement in this topic. We further discuss steepest ascent and related methods at a somewhat more advanced point of view in Chapter 6.

2.3 Problems

1 Consider the first order model $\widehat{y} = 100 + 5x_1 + 8x_2 - 3x_3$. This model was fit using an unreplicated 2^3 design in the coded variables $-1 < x_i < 1, i = 1, 2, 3$. The model fit was adequate and $S_b^2 = 2.54$. The region of exploration on the natural variables was

ξ_1 = temperature $(100, 110$ degrees C)

ξ_2 = time (1, 2 hrs.)

ξ_3 = pressure (50,75 psi)

a) Using x_2 as the variable to define a step size along the steepest ascent direction, choose a step size large enough to take you to the boundary of the experimental region in that particular direction. Find and show the coordinates of this point on the path of steepest ascent in the coded variables x_i.

b) Find and show the coordinates of this point on the path of steepest ascent from part a) using the natural variables.

c) Find a unit vector that defines the path of steepest ascent.

d) What step size multiplier for the unit vector in c) above would give the same point on the path of steepest ascent you found in parts a) and b)?

e) Find the fraction of directions excluded by the 95% confidence cone of steepest ascent.

2 Consider the first order model $\widehat{y} = 14.4 - 5.09x_1 - 13.2x_2 - 4x_3$, where y denotes the life of a tool, in minutes, as a function of 3 process variables. This model was fit using an unreplicated 2^3 design in the coded variables $-1 < x_i < 1, i = 1, 2, 3$. The model fit was adequate with $S_b^2 = 1.74$. The region of exploration on the natural variables was

ξ_1= speed of cut (650 to 800 units of surface per minute, sfm)

ξ_2 = cut feed (0.01 to 0.03 inches per revolution, ipr)

ξ_3 = depth of cut (0.05 to 0.20 in.)

and the current operating conditions are the center of this region.

a) For steepest ascent purposes, the Quality Engineer chooses to decrease the coded variable x_2 one unit from the origin. Find the coordinates of the resulting point on the path of steepest ascent in all other coded variables x_i.

b) Find and show the coordinates of this point on the path of steepest ascent from part a) using the natural variables.

c) Find a unit vector that defines the path of steepest ascent.

d) What step size multiplier for the unit vector in c) above would give the same point on the path of steepest ascent you found in parts a) and b)?

e) Find the fraction of directions excluded by the 95% confidence cone of steepest ascent.

3 Consider the first order model: $\widehat{y} = 52.1 - 3.1x_1 + 6.4x_2 - 1.25x_3$, where the variance of the parameter estimates, s_b^2, equals 0.4 computed based on 5 replications of the center point of the design. Does the point $\boldsymbol{x}' = (-0.9, 1.0, -0.3)$ generate a direction vector inside the 95% confidence cone of steepest ascent?

4 Consider the tool-life experimental data shown in Table 2.6 [44]. Two factors were varied in a replicated 2^2 with three center runs. The tool life of the tool was considered by the first occurrence of 0.015 inch uniform flank wear, 0.004 inch crater depth, 0.030 inch localized wear, or catastrophic failure.

Table 2.6. Data for problems 4 and 5

Speed (sfm)	Feed (ipr)	Tool life (minutes)	
		Grade A	Grade B
500	0.015	67.0	84.4
500	0.015	101.9	91.2
500	0.015	63.6	66.7
800	0.015	23.5	16.0
800	0.015	17.6	15.2
800	0.015	21.3	17.6
500	0.027	17.9	24.6
500	0.027	25.3	15.3
500	0.027	25.4	30.4
800	0.027	0.4	1.1
800	0.027	0.6	0.5
800	0.027	0.5	0.9
650	0.021	21.4	11.8
650	0.021	19.2	8.9
650	0.021	22.6	10.6

a) Fit the best polynomial model you can to the "grade A" tool data. (Hint: transformations of the response may be necessary).

b) Find the direction of steepest ascent, and determine a point that would take the process to a distance approximately equal to the boundary of the current experimental region.

5 Consider the "grade B" tool life data in Table 2.6.

a) Fit the best polynomial model you can.

b) Determine a point on the direction of steepest ascent that would take the process to a distance approximately equal to the boundary of the current experimental region.

For the following problems, readers may wish to consult Appendix A.

6 Consider the following one factor experiment:

ξ_1	y (observed)
650	7
800	18
650	6
800	11
725	10

a) Using regression analysis, find the least squares estimates and the estimated variance of the parameters estimates in the model:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

where x_1 is a coded factor corresponding to ξ_1.

b) Find Cook's D_i diagnostic statistics based on the studentized residuals (r_i). Which point(s) seem to be outliers and/or influential?

c) Test the significance of the regression model. What are the null and the alternative hypotheses?

d) Test for lack of fit of the linear model. What are the null hypothesis and the alternative hypotheses?

7 Assume a 1st order model is to be fit in k factors. From $\partial R(\boldsymbol{\beta})/\partial\beta_j = 0,\ j = 0, 1, 2, \ldots, k$, find the set of normal equations. Do not use any matrix notation.

8 Consider the following "unit length" coding convention:

$$x_{iu} = \frac{\xi_{iu} - \overline{\xi}_i}{s_i}, \quad i = 1, \ldots, k;\ u = 1, 2, \ldots, n$$

where

$$\overline{\xi}_i = \frac{\sum_{u=1}^{n} \xi_{iu}}{n}, \quad i = 1, \ldots, k$$

and

$$s_i = \sqrt{\sum_{u=1}^{n} (\xi_{iu} - \overline{\xi}_i)^2}, \quad i = 1, \ldots, k.$$

Show that the $\boldsymbol{X}'\boldsymbol{X}$ matrix will contain the correlation matrix of the controllable factors. Assume a first order model for simplicity.

9 It is well-known in regression analysis that

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$$

a) Interpret this expression using the concept of "projections" described in Appendix A. Draw a picture of the corresponding vectors to explain.

b) Does this expression imply that $\boldsymbol{Y} = \boldsymbol{\varepsilon}$? Explain.

10 Assume we fit a linear model with an intercept. Argue why $\mathbf{1}'\boldsymbol{H} = \mathbf{1}'$ where $\mathbf{1}$ is an $n \times 1$ vector of ones. What does this imply for the average prediction, $\overline{\widehat{Y}}$ and the average response $\overline{Y}$? Is $\mathbf{1}'\boldsymbol{H} = \mathbf{1}'$ if the model has no intercept? Why?

11 Consider the basic ANOVA for a linear regression model. Show that the following expressions for the "corrected sums of squares" are true. Do *not* use the relation $\text{SS}_{total} = SS_{reg} + SS_{error}$:

a) $\text{SS}_{total} = (\boldsymbol{Y} - \overline{Y}\mathbf{1})'(\boldsymbol{Y} - \overline{Y}\mathbf{1}) = \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2$;

b) $\text{SS}_{regression} = (\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1})'(\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1}) = \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - n\overline{Y}^2$. (Hint: use the result in Problem 10).

12 Show that $\overline{h} = p/n$. (Hint: use the fact that $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$, where "tr" means the trace; see Appendix C.). Relate this fact to

$$\sum_{i=1}^{n} \frac{\text{Var}\ \widehat{y}(\boldsymbol{x}_i)}{\sigma^2} = p.$$

What is this last expression saying in favor of simpler models (i.e., models with fewer parameters)?

13 Show that a random variable $W \sim F_{1,n-p}$ is equivalent to $W \sim t^2_{n-p}$.

Chapter 3

EXPERIMENTAL DESIGNS FOR FIRST ORDER MODELS

It appears to me, Sancho, that there is no saying that is not true, because they are all sentences taken from experience, mother of all sciences.

—Miguel de Cervantes Saavedra (1547–1616)

An obvious choice of experimental design for fitting a first order model is the class of 2-level factorial designs. Within this class, we have full 2^k designs, that is, factorial designs in which all k factors are varied only at two levels (a "low" and a "high" level, coded into the $(-1,1)$ convention discussed in Chapter 1) and all 2^k combinations are tried, hence their name. These experimental designs actually allow to estimate the more complicated model

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \sum_{j} \beta_{ij} x_i x_j + \sum_{i<j} \sum_{j<l} \sum_{l} \beta_{ijl} x_i x_j x_l + \cdots + \beta_{123\ldots k} x_1 x_2 \cdots x_k + \varepsilon \tag{3.1}$$

which contains all interactions up to order k. An alternative to this kind of design is to run instead a fraction of a full 2^k design and then either assume higher order interactions are negligible or eliminate negligible effects.

If a fractional factorial strategy is followed, a $1/r$ fraction (with r a multiple of 2, i.e., one half, one quarter, etc.) is selected from a 2^k. Denoted by 2^{k-r} designs since that is the number of factor combinations they have, these designs have been studied for a long time. They save runs ($2^{k-r} < 2^k$, evidently), but there is a price to pay. If the true model contains interactions as in the model

shown above and a 2^{k-r} design is run, it will result in some biased parameter estimates, the biases depending on the particular fraction selected.

In this chapter, we consider variance and bias properties of designs used to fit a first order model, i.e., a model with *no interactions*. We call these, by extension, *first order designs*. We concentrate mainly on 2^k, 2^{k-r} and related 2-level designs, such as Plackett-Burman designs.

3.1 Variance Properties of 2-level Factorial Designs for a First Order Model

In this section we will prove the following statement:

THEOREM 3.1 *Consider a process which follows the first order model*[1]

$$y_j = \beta_0 x_{0j} + \sum_{i=1}^{k} \beta_i x_{ij} + \varepsilon_j, \quad j = 1, \ldots, N \tag{3.2}$$

and an associated $N \times (k+1)$ *design matrix* $\boldsymbol{X} = [x_{ij}]$ *such that each of its columns* i *must satisfy the constraint* $\sum_{j=1}^{n} x_{ij}^2 \leq c_i,\ i = 0, \ldots, k$. *Then if* $\boldsymbol{X}$ *is:*

a) *orthogonal and*

b) *such that* $\sum_{j=1}^{N} x_{ij}^2 = c_i,\ i = 0, \ldots k,$

then it minimizes Var$(b_i)/\sigma^2, i = 1, 2, \ldots, k,$ *for the given number of runs* N.

A few remarks are in order before showing the truth of this statement:

- A limit in the sum of squares of each column is needed since otherwise the elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ could decrease indefinitely by making the levels of the factors (elements of $\boldsymbol{X}$) increase indefinitely. The sum of squares limitation can be seen as defining the maximum *size* of the DOE on the factor space (the larger the size, the lower the variances of the parameter estimates of a first order model). Usually, this will be given by some region of interest or operability. An instance of such restriction are cuboidal designs where the design levels are constrained by $-1 \leq x_{ij} \leq 1$. In such case we clearly have that $c_i = N$. This implies that a 2-level design that places the two levels at the extreme values are variance optimal for a first order model.

[1] Usually, $x_{0j} = 1$ if the model has an intercept.

- Based on this theorem we will use the term *first order variance optimal design*, but this means optimal given the design size N, since $c_i = \sum_{i=j}^{N} x_{ij}^2$, which clearly depends on N. It will be seen in Section 5.7.3 that first order designs that satisfy conditions a) and b) of the theorem are *D-optimal*, an important design optimality notion which will be introduced in Chapter 5.
- Note the theorem implies that 2-level factorial designs with center points (with the constraint $-1 \leq x_{ij} \leq 1$) are not variance optimal.
- The first column of $\boldsymbol{X}$, is set equal to $\mathbf{1}$ (a $N \times 1$ vector of ones) if the model contains an intercept, thus $x_{0,j} = 1, j = 1 \ldots, N$. This column always satisfies the sum of squares constraint (for $c_0 = N$) and all other columns need to be set orthogonal to the first one (and to each other).
- Finally, an important aspect to notice is that the theorem relates to the *variance properties* of the parameter estimates, not to their *bias* properties.

Many authors have offered proofs of this theorem, some apparently incomplete [16, 112] or too elaborated [76]. Plackett and Burman [128] proved this theorem for the case when the levels of the factors are constrained to $-1 \leq x_{ij} \leq 1$ (so $c_i = N$), concluding that their renowned designs are variance optimal for a first order model if $-1 \leq x_{ij} \leq 1$. The following is perhaps the simplest proof around, based on triangularization of a symmetric, positive definite matrix, and is due to Tocher [150].

Proof. Write the model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{X}$ is $N \times p$ and consider the matrix $\boldsymbol{X}'\boldsymbol{X}$. Given the stated constraint, the diagonal elements of the $\boldsymbol{X}'\boldsymbol{X}$ matrix can at most be equal to c_i, and the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ can be no smaller than $1/c_i$. The $\boldsymbol{X}'\boldsymbol{X}$ matrix is symmetric and positive definite, therefore a non-singular, upper triangular[2] matrix $\boldsymbol{U}$ can be found such that $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{X}'\boldsymbol{X}$. Furthermore, $\boldsymbol{V} = \boldsymbol{U}^{-1}$ is also upper triangular. Then we have that $v_{ii} = 1/u_{ii}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{V}\boldsymbol{V}'$. Consider the ith diagonal element of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix, call it α_i. We then have that

$$\alpha_i = \sum_{j=1}^{p} v_{ij}^2 \geq v_{ii}^2 = 1/u_{ii}^2 \geq 1/\sum_{j=1}^{p} u_{ji}^2 = 1/c_i.$$

For α_i to achieve its lower bound of $1/c_i$, both inequalities must become equalities, and for this to happen, we must have that $v_{ij} = 0$ for all entries in the

[2]That is, $u_{ij} = 0$ for $i > j$.

$\boldsymbol{V}$ matrix for which $i \neq j$. This means that $\boldsymbol{V}$ must be diagonal, and implies that $\boldsymbol{U}$ must be diagonal too. This finally implies that $\boldsymbol{X}'\boldsymbol{X}$ must be diagonal. Therefore, an orthogonal design for which $\sum_{j=1}^{N} x_{ij}^2 = c_i,\ i = 0, \ldots, k$ minimizes $\text{Var}(b_i)/\sigma^2 = [(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{ii},\ i = 0, \ldots, k$ for the given N. ■

Despite not resulting in variance optimal designs, it is customary to add replicated center runs to a 2-level factorial in order to be able to test for curvature and test for lack of fit. Furthermore, the center point provides a model-independent estimate of the error variance, and, if the number of center point replicates to total number of runs is small, the variance of the parameter estimates increases only a little over its minimum.

If the design has center points but is orthogonal, then still some advantages are kept. In particular, the parameter estimates are uncorrelated, so tests on their significance are independent. Useful first order orthogonal designs include:

- 2^k and 2_R^{k-r} $(R \geq III)$ designs[3];
- Simplex designs;
- Plackett Burman (PB) designs.

Simplex and PB designs are saturated for a first order model in k factors, i.e., $n = p$, so some replicates are necessary if statistical tests are desired. Note that $2^k, 2^{k-r}$ without center points, Simplex and PB designs are variance optimal according to Theorem 3.1 under the constraint $c_i = N$, although good practice in these designs is to add center points as mentioned before[4]. We treat PB designs in more detail in Section 3.4.

We finally point out in this section that the optimality of a design according to Theorem 3.1 relates to the variance of the parameter estimates, but having low variances in the parameter estimates translates in more precise response predictions, since the prediction variance will also be low. For a first order model fit with an orthogonal design according to the theorem, we have that, for a point in the DOE used to fit the model

$$\text{Var}(\widehat{y}(\boldsymbol{x})) = \sigma^2 \boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$$

[3]The subscript R denotes the *resolution* of a factorial design, a term we define in Section 3.3.

[4]The Simplex designs do not satisfy $-1 \leq x_{ij} \leq 1$ but satisfy the stated sum of squares constraint for $c_i = N$, hence they are also optimal.

where $\boldsymbol{x}'_m = f(\boldsymbol{x})' = (1, \boldsymbol{x}') = (1, x_1, x_2, \ldots, x_k)$ (we say that $\boldsymbol{x}_m$ is $\boldsymbol{x}$ in "model form") will result in

$$\text{Var}(\widehat{y}(\boldsymbol{x}))/\sigma^2 = 1 + \rho^2$$

where $\rho^2 = \boldsymbol{x}'\boldsymbol{x}$. Notice how the variance of the prediction at a point $\boldsymbol{x}$ is not a function of where the point is but only of the distance of the point with respect to the origin, as given by the quantity ρ. This indicates that all points separated a distance of ρ from the origin (i.e., points on a circle with radius ρ) will be predicted equally well. Later on we will refer to this variance property of a design as **rotatability**.

3.2 Useful 2-level First Order Designs

In this section we look in more detail at 2-level designs, paying particular attention to the bias aspects of these designs. The literature on DOE uses terminology that has an origin in the type of Agricultural experiments from where designed experiments originated, based mainly on the work by Fisher and Yates. We will explain some of this terminology in what follows.

3.2.1 2^k Designs

A 2^k design consists of running all combinations of runs at the ± 1 levels of the coded factors. Graphically, the design points correspond to the corners of a k-dimensional hypercube. Table 3.1 shows an instance of an *unreplicated* 2^4. As k gets larger, these designs get more impractical due to their large number of runs, even if no replications are conducted. The labels in the columns of Table 3.1 are one usual way of denoting the factors in the DOE literature[5]. The two levels of each factor are simply denoted by $+1$ and -1. In the DOE literature, when non-quantitative factors are considered it is customary to use the symbols $+$ and $-$, but this will not be necessary in this book as we assume continuous factors.

Construction of 2^k designs is easy: the so-called **standard order** shown in the table is used. This consists of writing 2^{j-1} cells with a "-1" followed by 2^{j-1} cells with a "$+1$" in each column j, $j = 1, \ldots, k$. The first row has all entries equal to -1.

[5]Another convention is to use numbers to denote the columns.

Table 3.1. A design matrix ($\boldsymbol{D}$) for a 2^4 experimental design

A	B	C	D
-1	-1	-1	-1
1	-1	-1	-1
-1	1	-1	-1
1	1	-1	-1
-1	-1	1	-1
1	-1	1	-1
-1	1	1	-1
1	1	1	-1
-1	-1	-1	1
1	-1	-1	1
-1	1	-1	1
1	1	-1	1
-1	-1	1	1
1	-1	1	1
-1	1	1	1
1	1	1	1

The total number of runs in a 2^k design is $N = n2^k$, where n is the number of times we replicate the DOE. Obtaining the effect estimates and the sum of squares necessary to do an ANOVA and test for significance of the different effects, etc., can be done using the ordinary least squares (OLS) formula with:

$$\boldsymbol{X} = [\mathbf{1}|\boldsymbol{D}|\text{other columns}]$$

where the "other columns" are columns that result from the element by element multiplication of the columns in $\boldsymbol{D}$ and that relate to each of the l-factor interaction terms in the model.

An elegant way of constructing the $\boldsymbol{X}$ matrix of a 2^k design for model (3.1) is to define the basic matrix[6]

$$H_2 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

[6] H_2 is a normalized Hadamard matrix of order 2, and the procedure indicated generates normalized Hadamard matrices of order N where $N = 0 (\text{mod } 4)$ [69].

Then the 2^k matrix $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \underbrace{H_2 \otimes H_2 \otimes H_2 \cdots \otimes H_2}_{k \text{ times}}$$

where $\otimes$ denotes the Kroenecker direct product (see Appendix C).

Example. For a 2^3 matrix, the $\boldsymbol{X}$ matrix for model (3.1) is

$$\boldsymbol{X} = H_2 \otimes H_2 \otimes H_2 = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Note how the columns associated with effects A, B, and C are columns 2, 3, and 5, as column 4 is the AB interaction, and so on. ■

Historically, the computations based on inverting the $\boldsymbol{X}'\boldsymbol{X}$ matrix were avoided in the pre-computer era thanks to using 2-level factorials. The manual computations are based on the concept of *treatment totals* and *contrasts*. We now review them briefly. For example, for a 2^2 design, we label the four "treatments" (i.e., combinations of factors) as $(1), a, b$, and ab, when writing the design in standard order. It is customary to let these same lowercase letters also represent the combination totals, i.e., the sum of the observed responses when that particular combination was run. See Table 3.2. Note how the combination a corresponds to the run where factor A is at its high level and all other factors are at their low level, and similarly for combination b. The first

Table 3.2. A 2^2 design showing the combination totals

Treatment combination	A	B	$y_{\text{1st rep.}}$	$y_{\text{2nd rep.}}$	Totals
(1)	-1	-1	3.1	5.4	8.5
a	1	-1	10.2	8.2	18.4
b	-1	1	11.6	9.1	20.7
ab	1	1	8.9	7.7	16.6

combination, denoted by a (1), contains all factors at their low level. If the first observed response at these settings is 3.1 and the second one is 5.4, the total is 8.5, so we write $(1) = 8.5$. A **contrast** for a particular column is the element by element product of one of the columns in $\boldsymbol{D}$ and the column of totals. Thus, using DOE notation we have that

$$\text{Contrast}_A = -(1) + a - b + ab = -8.5 + 18.4 - 20.7 + 16.6 = 5.8.$$

The contrasts compare (or contrast, hence its name) the runs where factor A is at its "high" level (a and ab) with the runs where it was set at its "low" setting ((1) and b).

The **effect** (the main effect, in this case) of factor A is given by $\text{Contrast}_A/4 = 1.45$. In general, for any 2^k or 2^{k-r} experiment[7] the effect of any factor or factor interaction is given by

$$\text{Effect}_{ABC\cdots} = \frac{\text{Contrast}_{ABC\cdots}}{N/2} \tag{3.3}$$

where N is the *total* number of runs. The effect for a 2-factor interaction is obtained by first obtaining the corresponding column from the basic DOE and then computing the contrast. For example, for the AB interaction in the 2^2 design just shown, we have that $AB = (1, -1, -1, 1)'$ which gives a contrast of -14 so $\text{Effect}_{AB} = -3.5$. The usual notation for effects in the DOE literature is to use the letters denoting the columns to also denote the corresponding effect. Thus, in the 2^2 design we have that $A = 1.45$ and $AB = -3.5$.

The relation between the effects as computed using (3.3) and the $\widehat{\beta}_1$ OLS estimate is that (3.3) measures the *average response as we change the factor from low to high*, which implies a net change in the regressor of 2 units. The model parameters β_i measure the average change in the response as we change the associated factor from 0 to 1, *half the range* than when using (3.3). Thus, $\widehat{\beta}_i = \text{Effect}_{x_i}/2$.

The sums of squares in a 2^k or a 2^{k-r} design can also be computed based on the contrasts according to

$$SS_{ABC\cdots} = \frac{\text{Contrast}^2_{ABC\cdots}}{N}.$$

Thus, in the 2^2 example we have that $SS_A = 4.205$, $SS_B = 13.52$, and $SS_{AB} = 24.5$. Each of these sum of squares has a single degree of freedom

[7] This is true in general for any balanced 2-level design.

Table 3.3. ANOVA for a 2^k design, showing all the $2^k - 1$ effects (m.e.=main effect, f.i.=factor interaction)

	Source	dof	SS	MS	F_o
	A	1	SS_A	$SS_A/1$	$\frac{MS_A}{MS_{error}}$
	B	1	SS_B	$SS_B/1$	$\vdots$
$\binom{k}{1} = k$ m.e.	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	K	1	SS_K	$SS_K/1$	
	AB	1	SS_{AB}	$SS_{AB}/1$	
$\binom{k}{2}$ 2-f.i.	AC	1	SS_{AC}	$SS_{AC}/1$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	ABC	1	SS_{ABC}	$SS_{ABC}/1$	
$\binom{k}{3}$ 3-f.i.	ABD	1	SS_{ABD}	$SS_{ABD}/1$	
$\vdots$	$\vdots$	1	$\vdots$	$\vdots$	
$\binom{k}{k} = 1$ k-f.i.	$ABC\ldots K$	1	$SS_{ABC\ldots K}$	$SS_{ABC\ldots K}/1$	
	Error	$2^k(n-1)$	SS_{error}	$SS_{error}/2^k(n-1)$	
	Total	$n2^k - 1$	SS_{total}	SS_{total}	

(we are essentially squaring the difference between two average responses, one obtained when the factor is high and one when the factor is low). There is a total of 2^k parameters in model (3.1), so the 2^k design is saturated for such model. The ANOVA table for a 2^k design and model (3.1) is shown in Table 3.3.

If it is desired to estimate model (3.1), note that replicates (i.e., $n > 1$) are necessary to perform the ANOVA and determine the significance of the effects (the effects themselves can be estimated, however). This is because the degrees of freedom for error are $2^k(n-1)$. Alternatives to replicating the whole design are as follows:

1 Do not replicate the design and assume higher order interactions (say, third order and higher) to be negligible and fit a model for all main effects and 2 factor interactions. The rationale is that high order interactions are usually not significant and are difficult to interpret in most cases. The neglected interactions are then "pooled" (added up) into SS_{error};

2 Do not replicate the design but compute all effects. Plot all the effects on a normal probability plot (NPP). Those effects that are insignificant will tend to be distributed along a line on the graph which corresponds to a zero mean, normally distributed effects (they are normal if the errors are normal). Pool all the negligible effects according to the NPP into SS_{error}. This and the previous option are based on the *Sparsity of Effects Principle* introduced in Chapter 1;

3 Replicate only the center points. This allows to test for curvature and for lack of fit, and provides a model independent estimate of σ^2. We discuss the addition of center points in Section 3.8;

4 Run a 2^{k-r} fractional factorial or a Plackett-Burman design.

In the next two sections, we discuss Fractional Factorials and Plackett-Burman designs.

3.3 2^{k-r} Designs. Resolution of a Fractional Factorial. Construction of 2^{k-r} Designs

A 2^{k-r} fractional factorial has the advantage of requiring less runs than a 2^k but there is a price to pay: if higher order interactions exist, the parameter estimates will be biased, with the bias depending on the actual DOE used and the real function the process follows.

Replicating a 2^{k-p} design is not common since it defeats its main advantage, namely, saving runs, without helping to solve its main deficiency, that is (as we will see), the biased estimates they produce.

It has been customary in the DOE literature to analyze what would happen if we run a particular 2^{k-r} design when in fact the model (3.1) is the true process description. Suppose we have 4 factors and decide to run a half of the full 2^4 design, i.e., we want to run a 2^{4-1} design (8 runs) . Which half shall we select? We could choose the 8 runs from a full 2^4 design in $\begin{pmatrix} 16 \\ 8 \end{pmatrix} = 12870$ different

Table 3.4. A 2^4 design and its two 2^{4-1} halves determined according to the $ABCD$ (highest) interaction

I	A	B	C	D	AB	AC	AD	BC	BD	CD	ABC	$ABCD$
1	−1	−1	−1	−1	1	1	1	1	1	1	−1	1
1	1	−1	−1	1	−1	−1	1	1	−1	−1	1	1
1	−1	1	−1	1	−1	1	−1	−1	1	−1	1	1
1	1	1	−1	−1	1	−1	−1	−1	−1	1	−1	1
1	−1	−1	1	1	1	−1	−1	−1	−1	1	1	1
1	1	−1	1	−1	−1	1	−1	−1	1	−1	−1	1
1	−1	1	1	−1	−1	−1	1	1	−1	−1	−1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
1	−1	−1	−1	1	1	1	−1	1	−1	−1	−1	−1
1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	−1
1	−1	1	−1	−1	−1	1	1	−1	−1	1	1	−1
1	1	1	−1	1	1	−1	1	−1	1	−1	−1	−1
1	−1	−1	1	−1	1	−1	1	−1	1	−1	1	−1
1	1	−1	1	1	−1	1	1	−1	−1	1	−1	−1
1	−1	1	1	1	−1	−1	−1	1	1	1	−1	−1
1	1	1	1	−1	1	1	−1	1	−1	−1	1	−1

ways. Fortunately, this type of designs have been studied for over 80 years, and a huge literature exists on how to proceed.

Table 3.4 shows only *one* of the possible partitions we could choose, with the two halves of 8 runs listed one after the other. The columns shown are the basic design columns (labeled A to D), all the 2-factor interaction columns in addition to one 3-factor and one 4-factor interaction column (the first column, I, is a column of all ones, which will be discussed shortly). These particular selection of halves was obtained as follows. The A, B, and C columns were written in standard order. Notice how the D column is not in standard order, although we could re-arrange the rows and write them such that D is also in standard order. Instead of doing that, we arrange the top 8 runs (rows) such that their entries in the $ABCD$ column equal 1, and the bottom 8 runs are such that for them $ABCD = -1$. As a result of this re-arrangement of the rows, notice that for the first half (first 8 rows) we have that the numbers in column D

equal the numbers in the column labelled ABC. It is easy to check that column ABC is obtained by multiplying element by element the numbers in columns A, B and C. Thus, we will write $D = ABC$ for the first half. Similarly, for the second half it is easy to see that $D = -ABC$.

With this, in the first half we have that $I = ABCD$ and for the second half we have that $I = -ABCD$, where I is a column of all ones. These expressions are the **defining relations** of each half, because with them we can construct the design. As it turns out, the algebra of columns in a 2-level factorial is not the standard algebra we learn in elementary school. There are two basic operations the algebra of columns obeys. Let W be any column of the design and let $(\cdot)$ denote element by element multiplication of two columns. Then:

$$W \cdot W = I$$

and

$$W \cdot I = W \text{ (i.e., get the same column).}$$

Thus, for example, if W is column A then $A \cdot A = I$, as it can readily be verified how by multiplying element wise any column in the design times itself gives a column of ones[8]. Thus the "times" operation of columns is not standard algebra. In the algebra of DOE columns, *the inverse equals the square*, with the column I playing the role of the identity[9]. Note how these rules work in the example: $I = ABCD$ implies $DI = ABCD^2$ which implies $D = ABC$.

So, to construct a 2^{4-1} design where $I = ABCD$ all we have to do is to create a table where columns A, B, and C are in standard order and add a column for D such that $D = ABC$. This will pick the top half in Table 3.4.

A central concept in fractional factorial designs is **resolution**. In the first half of Table 3.4 we have that $D = ABC$. This means that the contrast of column D and the contrast of column ABC are identical, so when estimating the effect of column D *we are also estimating the effect of column* ABC. We will say that effect D is **aliased** with effect ABC, since the same effect has "two names", or that the two effects are **confounded**. Using the algebraic properties of columns, it is not difficult to see some other aliased columns:

[8]Note how the powers are reduced modulo 2, i.e., $A^2 = A^4 = \ldots = A$, or $A^l = A^{[l(\text{mod } 2)]+1}$.
[9]Contrast with regular algebra where for any $x \neq 0$, $x^{-1}x = 1$ and number one is the identity element.

$$\begin{aligned} A &= BCD \\ B &= ACD \\ C &= ABD \\ AB &= CD \\ AC &= BD \\ AD &= BC \end{aligned}$$

The first 3 are aliases between a main effect (i.e., a single letter effect) and a 3 factor interaction not involving the letter. The last 3 aliases are between different 2-factor interactions. Thus, in this design, main effects are aliased with 3-factor interactions, and 2-factor interactions are aliased with one other 2-factor interaction. A design with this type of **confounding** is called a design of *resolution* IV (Roman four).

The term "resolution" was borrowed from optics. A design with higher resolution will allow us to "see" the effects more clearly. In the limit, a full 2^k design has full resolution to estimate in an unbiased manner all effects in model (3.1). In the following discussion, it is important to keep in mind that the importance of the effects in practice usually decreases with the number of factors in the effect. Thus, main effects (i.e., single factor effects) are usually more important that 2-factor interactions which in turn are usually more important than 3-factor interactions, etc. By "importance" we mean larger in absolute value. A more precise definition of resolution is as follows:

DEFINITION 3.1 *A design of resolution R is one in which no m-factor interaction effect is aliased with any other effect with less than $R - m$ factors.*

Therefore:

- A design of resolution $R = III$ has main effects (i.e., single letter effects) confounded with at least one 2-factor interaction, but main effects are not aliased with other main effects. For example, consider the 2^{3-1} design with $C = AB$, denoted 2^{3-1}_{III}.

- A design of resolution IV has main effects confounded with 3-factor interactions, and 2 factor interactions confounded with other 2-factor interactions. This is clearly better than a $R = III$ design since main effects are not

aliased with 2-factor interactions but with 3-factor interactions, which tend to be less important. The 2^{4-1}_{IV} design with $D = ABC$ shown earlier is an example.

- A design of resolution V has main effects confounded with 3-factor interactions, and 2-factor interactions are confounded with 3-factor interactions. An instance is a 2^{5-1}_{V} with $E = ABCD$. Note how this is a better design than a resolution IV design since 2-factor interaction effects are not aliased with other 2-factor interaction effects.

As a further analogy with Optics we will say that if a main effect or a 2-factor interaction is not aliased with any other main effect or 2-factor interaction, then these effects are **clear**. Thus, in a $R = IV$ design main effects are clear but 2-factor interactions are not, whereas in a $R = V$ design both main effects and 2-factor interaction effects are clear.

Higher order resolution designs are of course possible and desirable. Resolution III designs are useful for factor screening (see Chapter 1), that is, to discern from among a very large number of potential factors k which ones are the "active" factors in the process. If k is very large, it is prudent as a first approach to consider main effects only in a small DOE and eliminate inactive factors. The results of such experiment would then indicate what further experiments are necessary in the active factors, and in those experiments we should then use higher resolution. Resolution II designs are never considered because the main effects would be confounded with other main effects.

Consider the defining relations of the three DOEs listed above: $I = ABC$, $I = ABCD$ and $I = ABCDE$. The *words* on the right hand side of these expressions have lengths equal to the resolution to each design. Thus we have an alternative definition of resolution.

DEFINITION 3.2 *The resolution of a design is the shortest word in the complete defining relation.*

In the examples above, which were all half fractions $(r = 1)$, there was only one such word, so automatically it was the shortest.

Any 2^{k-r} design will have r **generators** and 2^{k-r} **alias chains** each consisting of 2^r elements. The generators are the words used to define the r columns when constructing a fraction. Products of these generators will give more

words, and the shortest of them gives the resolution of the design. An example shall illustrate these definitions.

Example. Consider a 2^{6-3} experiment with generators $D = AB$, $E = AC$ and $F = BC$. The defining relation is given by

$$I = ABD = ACE = BCF = DEF = ABEF = ACDF$$

where the first $3(= r)$ words are the generators, the next word is the product of all 3 generators, and the last 3 words are obtained as products of 2 generators at a time. This is the *complete* defining relation of the design. Its shortest word has 3 letters (I is not considered a word), hence $R = III$. The $8(= 2^{6-3})$ alias chains each of $2^r = 8$ elements give all the aliases of each effect, and are obtained from the complete defining relation by multiplying all terms by the effect of interest:

$$\begin{aligned}
I &= ABD = ACE = BCF = DEF = ABEF = ACDF = BCDE \\
A &= BD = CE = BEF = CDF = ABCF = ADEF = ABCDE \\
B &= AD = CF = AEF = CDE = ABCE = BDEF = ABCDF \\
C &= AE = BF = ADF = BDE = ABCD = CDEF = ABCEF \\
D &= AB = EF = ACF = BCE = ACDE = BCDF = ABDEF \\
E &= AC = DF = ABF = BCD = ABDE = BCEF = ACDEF \\
F &= BC = DE = ABE = ACD = ABDF = ACEF = BCDEF \\
AF &= BE = CD = ABC = ADE = BDF = CEF = ABCDEF
\end{aligned}$$

For instance, multiplying all words in the complete defining relation times A, we get:

$$A = BD = CE = BEF = CDF = ABCF = ADEF = ABCDE$$

which is the second alias chain listed above (the other 6 chains are obtained similarly). Thus, we see that while it is true that the main effect A is confounded with at least one 2-factor interaction, and hence this is a resolution III design, main effect A has many other aliases, including two 2-factor interactions, two 3-factor interactions, 2 four factor interactions, and one 5-factor interaction. ■

To construct a 2^{k-r} design, we would choose a design that has the highest resolution possible for the experimental budget (usually, a fixed number of

runs) we have available. Catalogs of these designs exist that give generators that guarantee maximum resolution for the given sample size. An on-line catalog is available from the NIST/Sematech Handbook of Engineering Statistics (Section 5.3.3.4.7), a public document available at:

http://www.itl.nist.gov/div898/handbook/index.htm.

These DOE catalogs give, for a given number of factors k and runs N, the highest resolution possible that has **minimum aberration**. In general, to achieve higher resolution, one needs larger designs, up to having full resolution for all terms in model (3.1) when using a full 2^k design. But as it turns out, there might be several 2^{k-r}_R designs for a particular choice of k, r, and R, since they all have at least one word of length R. The minimum aberration design "breaks the ties" and chooses the design with fewer words of shortest length. This results in simpler confounding (simpler alias structure) among the designs tied according to the resolution criterion. See [157] for more details.

Given the r generators, to construct the design we write the first $k-r$ columns in standard order and the next r columns using each of the generators.

The resolution of a design can be increased from III to IV by "folding over" it (see Section 3.3.2 below). Likewise, particular alias chains can be broken and "clear" estimates can be obtained by adding more runs to the design. See [110, 157] for details.

The algebra of columns has a deeper mathematical foundation based on group theory. We show some of the basics of this foundation next.

3.3.1 Group Theory Aspects of 2^k and 2^{k-r} Factorials**

It turns out[10] that in a 2^k or 2^{k-r} design, the set of effects or columns in model (3.1), namely

$$\{I, A, B, \ldots, AB, AC, \ldots, ABCD\cdots K\}$$

and the corresponding set of treatments totals or rows

$$\{(1), a, b, \ldots, k, ab, \ldots, abcd\cdots k\}$$

constitute two instances of *Abelian groups*.

[10]This section has somewhat more advanced material and may be skipped on a first reading.

A *group* is a set G with a binary operation $*$ such that the operation:

1 is associative,

2 has an identity,

3 has an inverse operation.

If in addition, $*$ is commutative, then the group is said to be Abelian.

An example of an Abelian group is the integers with addition as the $*$ operation (note a group could be infinite in size). Readers may wish to consider whether the integers under multiplication form a group or not.

For the group of effects in 2^k or 2^{k-r} factorial, the group operation $*$ is defined as $A * B \doteq [A_i \cdot B_i]$ Thus, if $A = [-1, +1, -1, +1]'$ and $B = [-1, -1, +1, +1]'$, then $A * B = [+1, -1, -1, +1]'$ and we denote it (for short) as AB. This is *not* the dot product of the two vectors, although we will use the word "product" to refer to it. This operation does not obey the usual algebraic rules.

Notice all properties of Abelian groups are satisfied for the $*$ operation:

1 $A * B * C = (A * B) * C$

2 $A * I = A$

3 $A * A = I$ (the inverse operation consists of multiplying by itself, i.e., $A^{-1} = A$ (we write $A^2 = I$ for short)

4 $A * B = B * A$

For the set of treatment combinations $\{(1), a, b, \ldots, k, ab, \ldots, abcd \cdots k\}$ the operation $*$ that makes this set an Abelian group is more complicated. In this case, $a * b \doteq [((1)_i + a_i + b_i) \pmod 2]$ where $a_i = 0$ if there is a "-1" in the ith position of this *row* vector and $a_i = 1$ if there is a "+1" in the ith position (similarly for b_i. $(1)_i$ denotes the ith position in the row labeled (1)).

For example: consider a 2^3 design. Then $\{(1), a, b, c, ab, ac, bc, abc\}$ forms a group under the operation $*$ defined above. Take $a * b = [(0 + 1 + 0)mod\ 2, (0 + 0 + 1)mod\ 2, \ldots, (0 + 1 + 1)mod\ 2] = [1, 1, 0, 1, 0, 0, 0]$ so we get $ab = [+1 + 1 - 1 + 1 - 1 - 1 - 1]$.

The most enlightening property of groups that helps understanding of the algebraic properties of DOE columns is the fact that a group is *closed* under the

"*" operation. This means that any product of effects is another effect in the group. For example, consider a 2^{5-2} design with generators $E = AD, B = CE$. The generators imply that $I = ADE = BCE$. The product $ADE * BCE = ABCD$ makes up the complete defining relation:

$$I = ADE = BCE = ABCD$$

Any binary product of elements in the group $G = \{I, ADE, BCE, ABCD\}$ results in another element of the group. Thus, for instance

$$ABCDE * ADE = BCE \in G$$

or

$$ABCD * BCE = ADE \in G$$

$$ABCD * ADE * BCD = I \in G$$

thus the result is always in the group G. We see that the closeness property is a consequence of the way the operator * was defined.

In a full factorial, the closeness property of the group of effects holds too. Thus, in a 2^2 design, the group is $G = \{I, A, B, AB\}$ and any product of elements results in another element in G, e.g., $AB * A = B$, $A * I = A$, etc. The difference with a group of effects in a 2^{k-r} design is that in a fraction, the group of effects has members that are equal, and the alias structure tells us which are those elements. So, in a 2^{3-1} with $C = AB$ we know that the group of effects is $G = \{I, A, B, C, AB, BC, AC, ABC\}$ but then $A = BC$, $B = AC$, $C = AB$. The terms *defining relation* and *generator* of a design ($I = ABC$ and ABC, respectively, for the 2^{3-1} design) are terms borrowed from group theory, as they are the defining relation and the generator of the group G.

In a full factorial there is no such defining relation for the group, which simply contains all the effects as distinct elements.

3.3.2 Folding Over Resolution III Designs

It is always wise never to expend the experimental budget in a first experiment. It is always better to start with a small fractional factorial to which we can add more runs to increase the resolution or break alias chains that obscure model interpretation. One useful technique for doing this is **folding over** a resolution III design. Folding over consists in running a second DOE where one

or more columns in the design have their signs switched. If the initial design is a 2^{k-p}_{III}, this either breaks the aliases between a main effect and its two factor interactions or breaks the aliases between all main effects and the two factor interactions, respectively. The following summarizes the main results in each of these two cases.

1. Folding over by switching the sign in one column only of a 2^{k-p}_{III}

Here we switch the sign of a single column (a main effect or a two factor interaction. If l_A denotes the contrast for column A (this will also refer to the contrast of a two-factor interaction in a resolution III design, hence we label it by the letter l for *linear contrast*) then:

- we get the main effect of that factor (say, A), from $\frac{1}{2}(l_A + l_{A'})$;
- we get all the two-factor interactions that contain factor A from $\frac{1}{2}(l_i - l_{i'}),\ i \neq A$;
- The generators of the second fraction are the same as those in first fraction except that the words containing the "switched" factor (letter) are multiplied by -1;
- The "assembled" design (i.e., putting the 2 fractions together) has $U+L-1$ generators where L is the number of words of Like sign in each of the two fractions and U is the number of words of Unlike sign (note this means you get one generator less for the assembled design than for each of its two fractions);
- The complete defining relation of the assembled design is shortened. It is made up of the L generators of like sign and the $U-1$ generators *which are even products of independent words of unlike sign*.

Example. Consider a 2^{7-4} fraction with generators: ABD, ACE, BCF, and $ABCG$. If we switch column D, then the fold over fraction has generators: $-ABD$, ACE, BCF, $ABGG$. This implies that the assembled 2^{7-3} design has $L+U-1=3$ generators: $U-1=0$ words which are products of unlike sign words and the $L=3$ words of like sign: $ACE = BCF = ABCG$. The complete defining relation is $I = ACE = BCF = ABCG = ABEF = BCEG = AFG = CGEF$. By doing this, D and all its two factor interactions are "clear". However, *we still have a resolution III design.* ■

2. Switch signs in all columns of a 2^{k-p}_{III} ("full" fold over):

If we switch the signs of *all* the columns to create the second experiment (getting what some authors call a "reflection"), then:

- we get all main effects clear from $\frac{1}{2}(l_i + l_{i'})$;
- the two factor interactions will still be aliased (we get a resolution IV design). The chains of two factor interactions are estimated from $\frac{1}{2}(l_i - l_{i'})$;
- the generators of the second fraction are:
  - Same as those in first fraction if the number of letters in the generator is even;
  - Same as those in first fraction but with the sign changed if the number of letters is odd.
- The $U + L - 1$ generators of "assembled" design are the L words of like sign and the $U - 1$ generators which are independent even products of the U words of unlike sign.

Example. Suppose we initially run a 2^{7-4}_{III} fraction with generators: ABD, ACE, BCF, $ABCG$. Then the fold over obtained by switching the sign in all columns is a second 2^{7-4}_{III} fraction with generators: $-ABD, -ACE, -BCF$, $ABCG$. The assembled 2^{7-3}_{IV} design has generators $ABCG$ ($L = 1$ word) and $U - 1 = 2$ words that are independent even products of the $U = 3$ words of unlike sign[11], for example: $ABD * ACE = BCDE$ and $ABD * BCF = ACDF$. The complete defining relation of the assembled experiment is $I = ABCG = BCDE = ACDF = ADEG = BDFG = ABEF = CEFG$. ■

When doing a full fold over of a resolution III design, the words in the defining relation of the "assembled" design always have 4 or more letters, i.e., *we get resolution=IV for the combined design.* As proved by Margolin [92], this folding procedure is the only way to generate **minimal** resolution IV designs, defined as designs in k factors such as $N = 2k$. Not all folded designs of resolution III will result in minimal resolution IV designs, however, and in general we will get an assembled design with $N \geq 2k$.

[11] Any 2 such products, recall the set of words in the complete defining relation forms an Abelian group, see Section 3.3.1.

Folding 2^{k-p}_{IV} designs is also possible, but it generally results in assembled designs that have too many runs and is not recommended. Furthermore, the assembled design will *not* have resolution V, although it will have less aberration than the original design. A better approach is to add runs using the D-optimality principle (see Chapter 5).

3.3.3 Projection Properties of 2^{k-r} Designs

A projection of a design simply means to select some columns of the design and ignore the remaining ones. *Any 2^{k-r}_R design contains full factorials (possibly replicated) in every subset of $R-1$ variables*. We thus say that the **projectivity** of 2^{k-r}_R designs is $R-1$. This projection property of fractional factorial is useful in practice. Suppose in an experiment with k factors only $R-1$ are suspected to have a significant effect prior to running the experiment. We could then run a resolution R design. If our guess is correct and all these $R-1$ factors turn out to be the only significant ones after the experiment is conducted, we will end up with a full factorial on these effects, with which we can estimate all effects and interactions between them clear of any aliasing. This result gives an auxiliary criterion to aid selecting the minimum resolution we should require in a DOE.

More precisely, if we project a 2^{k-r}_R fraction into $R-1$ columns we get a full factorial since the shortest word in the defining relation has length R, thus choosing $R-1$ columns will not result in a alias relation between the selected columns. However, if we project the design into R or more factors, then whether or not the selected columns form a full factorial depends on whether or not the selected columns form a word in the defining relation. If they do, then the selected columns will form a replicated fractional factorial. If the do not, the selected columns will form a replicated full factorial.

Example. Consider a 2^{5-2} design with generators $D=AB$ and $E=AC$, shown in Table 3.5 and Figure 3.1. If we conduct the experiment and it turns out that factors D and E are not active, the remaining three factors A, B, C will form a full factorial (see Figure 3.2) because the word ABC is not a word of the complete defining relation given by $I=ABD=ACE=BCDE$ (see Figure 3.2). If in contrast, it turns out that C and E are not active, the remaining factors A, B, and D will form a fractional factorial, not a full factorial, because

Table 3.5. A 2^{5-2}_{III} design

A	B	C	$D = AB$	$E = AC$
−1	−1	−1	1	1
1	−1	−1	−1	−1
−1	1	−1	−1	1
1	1	−1	1	−1
−1	−1	1	1	−1
1	−1	1	−1	1
−1	1	1	−1	−1
1	1	1	1	1

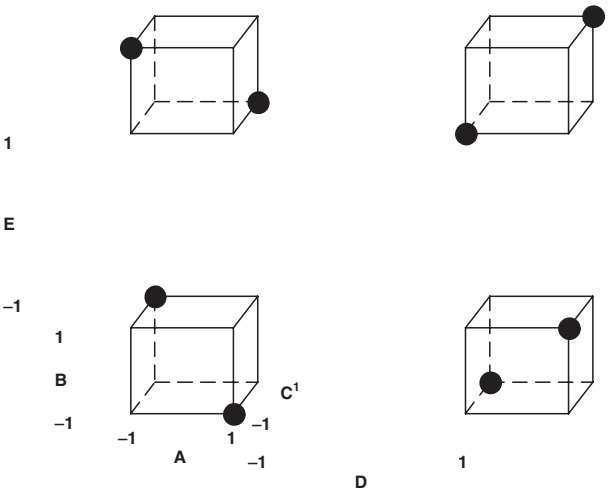


Figure 3.1. A 2^{5-2}_{III} design

the word ABD is in the defining relation, so for the remaining columns we clearly have $D = AB$ (the figure was created with the useful "Cube plot" option of Minitab, which does not show replicated points. The fractional factorial has actually two replicates in 3 different points). ■

3.4 Plackett-Burman Designs and Other Screening Designs

An economical class of 2-level designs was proposed by Plackett and Burman (PB) [128]. PB designs have resolution III and are saturated ($n = p$). They are constructed, in general, differently than 2^{k-r}_{III} designs. All PB designs have

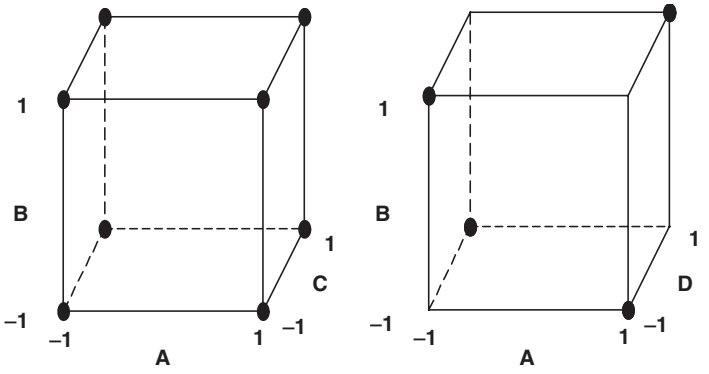


Figure 3.2. Projections of a 2^{5-2}_{III} design in A, B, and C (left) and on A, B, and D (right). The complete defining relation is $I = ABD = ACE = BCDE$

N a multiple of 4 and are orthogonal for a first order model (from Theorem 3.1, they are optimal for a first order model if we have the restriction $-1 \leq x_{ij} \leq 1$ which implies that $c_i = N$). If N is a power of 2, these designs are fractional factorials designs (these are sometimes are referred to as geometrical PB designs, and when $N \neq 2^k$ PB designs are called non-geometrical). Since these designs are mainly used for factor screening, not for optimization, we will discuss them only briefly.

PB designs have a more complicated and difficult to determine alias structure than 2^{k-r}_{III} designs. Each main effect is aliased with a long string of 2-factor interactions. The aliasing is not total but partial, in the sense that in a 2^{k-r}_{III} design main effects are identical to some 2-factor interactions, but in a PB design main effects are only *correlated* with the columns corresponding to 2-factor (and higher) interactions. Thus, in a PB design aliases are more gradual.

The attractiveness of the projection properties of PB designs has been subject to certain debate. PB designs for 12, 20 and 24 runs project into full factorials in any 3 factors, they therefore have projectivity 3. Resolution III fractional factorials, in contrast, have projectivity $R - 1 = 2$. Montgomery [110] indicates that for $N = 12$, a PB for 11 factors design projects into an unbalanced fraction in four factors (see leftmost display in Figure 3.3). Wu and Hamada [157] indicate that this fraction still has good efficiency to estimate main effects and 2-factor interactions among the remaining 4 factors if $k \geq 4$ and N not a multiple of 8 (which includes the $k = 11, N = 12$ case). They call this

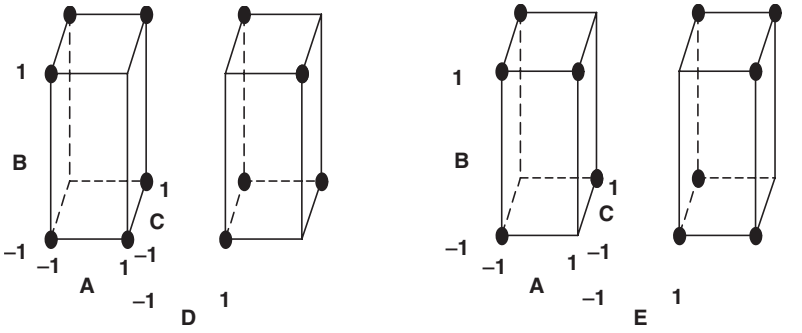


Figure 3.3. Projection of a PB design for $k = 11$ in $N = 12$ runs into factors A, B, C and D (left two cubes). On the right two cubes, projection of a 2^{11-7}_{III} design onto A, B, C, and E

property the *hidden projection property* of PB designs. Box et al. [25] indicate that it is the projection properties of these designs that makes them useful for factor screening.

Example. Consider a PB design for 11 factors in 12 runs. The design is shown in Table 3.6. If projected into every 3 columns, it results in a full factorial, but if projected into 4 columns (Figure 3.3, left panel) it results in a fractional factorial which is actually not balanced. An alternative to this design is to run a 2^{11-7}_{III} $(N = 16)$ with generators $E = ABC, F = BCD, G = ACD, H = ABD, J = ABCD, K = AB, L = AC$. If any three active factors form a word in the defining relation of the 2^{11-7} (the three-letter words in the defining relation are $ABK = ACL = AFJ = BEL = BGJ = CEK = CHJ = DEJ = DGL = DHK = FGK = FHL$), the resulting design will be a fraction of resolution III; otherwise the projection will be a full 2^3 factorial[12]. The right panel of Figure 3.3 shows the projection of this fractional factorial onto columns A, B, C, and E. Since $ABCE$ is clearly a word of the defining relation, we get a resolution IV fraction (this design when projected onto A, B, C, and D gives a full factorial in these factors, as they do not form a word in the defining relation). ■

Analysis of PB designs is difficult given the complex alias structure due to partial aliasing. Box et al. [25] recommend using Bayesian methods to

[12]In the absence of any *a priori* information about active factors, the probability of having 3 factors picked at random forming any one of the 12 words in the defining relation is $12/\begin{pmatrix} 12 \\ 3 \end{pmatrix} = 0.0545$.

Table 3.6. A Plackett-Burman design for $k = 11$ in $N = 12$ runs

A	B	C	D	E	F	G	H	J	K	L
1	−1	1	−1	−1	−1	1	1	1	−1	1
1	1	−1	1	−1	−1	−1	1	1	1	−1
−1	1	1	−1	1	−1	−1	−1	1	1	1
1	−1	1	1	−1	1	−1	−1	−1	1	1
1	1	−1	1	1	−1	1	−1	−1	−1	1
1	1	1	−1	1	1	−1	1	−1	−1	−1
−1	1	1	1	−1	1	1	−1	1	−1	−1
−1	−1	1	1	1	−1	1	1	−1	1	−1
−1	−1	−1	1	1	1	−1	1	1	−1	1
1	−1	−1	−1	1	1	1	−1	1	1	−1
−1	1	−1	−1	−1	1	1	1	−1	1	1
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

disentangle the complex alias relationships. Bayesian methods have also value in analyzing any other low-resolution design. They point out that a normal probability plots of effects, as used in 2^{k-r} designs, may be misleading due to the partial aliasing. See Part V for more on Bayesian methods.

Other screening designs apart of PB and 2^{k-r}_{III} designs are supersaturated designs, in which N is *less* than p, where the number of parameters of interest are all k main effects and all the 2-factor interactions. For the construction and analysis of these designs, see [85, 157].

3.5 Randomization and Blocking

Randomization allows to neutralize undesirable or uninteresting sources of error in an experiment that cannot be controlled. Running the experiments in random order is part of randomization. But not only that: in general, we should randomly assign the treatments ("design points") to the experimental units (see Chapter 1). If the experimental units differ, this will be balanced or averaged by the randomization step. It is always recommended that the experiments used to optimize a process be randomized. Here it is important that to truly have a randomized experiment, we should **reset** the factor settings before each new

run. Thus, suppose we have a factor that is difficult to vary, say temperature. Suppose we obtained the following test conditions after shuffling a DOE randomly: $100^\circ F, 150^\circ F, 150^\circ F, 100^\circ F$. Experimenters will likely *not* reset the temperature from run 2 to run 3, since the DOE calls for keeping the level at $150^\circ F$. However, resetting the level to some base value (in this case, waiting for some furnace to cool down, for example) and then increasing the temperature to $150^\circ F$ again is *exactly* what randomization calls for. If no resetting takes place, we call the DOE a design run in random order. If we reset, we call it fully randomized. Not resetting may induce what is called a "Split-Plot" structure[13] in the experiment, and the data analysis should take this into consideration [59].

The basis of statistical inference is that the probability model that generates the data is known and available. Therefore, it is necessary to generate the observations in such a way that the probability distribution of the various outcomes can be specified to the extent necessary to conduct the tests of hypothesis and compute the confidence intervals necessary in the analysis. This is exactly achieved by the randomization principle [134].

If a suspected source of variability can be controlled in the sense that we can vary the levels of the nuisance factor that creates the variability during the experiment, then **blocking** is an alternative technique that can be used in experimental design. For example, if we suspect that different batches of a raw material may induce variability in the response of interest, say, some quality characteristic of a manufactured product, and we are *not* interested in the effect of the batches, we can assign the experimental units to different batches of raw material, i.e., we can control which experimental runs are performed with which batches. Thus, we can form groups of experimental runs that use more homogeneous material (e.g., runs that use the same raw material). This will enhance the sensitivity of the experiment to detect *other* significant effects due to factor we are interested. Another instance of a blocking factor is simply the time in which experiments are conducted (e.g., different days). If we suspect or know that experiments conducted on Monday will give different results than those conducted on Friday, and we are not interested in the "day" factor, we can form blocks of experimental runs to separate this source of variability and reduce the experimental error. Blocking an effect reduces the experimental error since the source of variability of the nuisance factor is eliminated from it.

[13] See Section 9.4.

When blocking, we randomize the runs within each block, but not from block to block. We will discuss blocking only when we look at Central Composite and other second order designs in Chapter 6 and at Split Plot designs in Chapter 9. For blocking in factorial designs see [25, 110].

3.6 Designs for Models with Main Effects and 2-Factor Interactions

A similar result to Theorem 3.1 exists for a model with main effects and 2-factor interactions:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \sum_{j} \beta_{ij} x_i x_j + \varepsilon.$$

Thus, an orthogonal design that places the levels of the factors at the extremes minimizes the variances of the parameter estimates. Two instances of such orthogonal designs are 2^k designs and 2^{k-r} fractional factorials with resolution $\geq V$. A fraction with resolution lower than five will result in an $\boldsymbol{X}'\boldsymbol{X}$ matrix that is non-invertible.

3.7 Bias Properties of First Order Designs

The defining relation gives the alias structure of a 2^{k-r} design if the model is (3.1). These aliases say that parameter estimates will be biased. If the model is more complex than (3.1), the defining relation method does not work and we need to find the aliases some other way. A useful concept to study the potential bias of parameter estimates with respect to more general "true" models is the *Alias matrix*. Suppose our assumed model is

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

of order d_1, with OLS estimate $\widehat{\boldsymbol{\beta}_1} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}$. Suppose the true model the process follows is

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

of order $d_2 > d_1$. The bias of the parameter estimates in our model will then be obtained from

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}_1}] &= E[(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}] = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{A}\boldsymbol{\beta}_2. \end{aligned}$$

Thus the bias in the parameter estimates is $\boldsymbol{A}\boldsymbol{\beta}_2$, where we call $\boldsymbol{A}$ the "alias" matrix, since it gives the relationships between each parameter in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. These are the aliases or confounding scheme of the DOE. Note how the biases depend on things we can control (the fitted model and the DOE used) and on something we cannot control (the true process model).

The bias in the parameter estimates is of concern in process optimization, since these biases are transmitted to the fitted function itself, just as the variance properties:

$$\text{Bias}(\widehat{\boldsymbol{Y}}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_1\boldsymbol{A}\boldsymbol{\beta}_2 - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2 = (\boldsymbol{X}_1\boldsymbol{A} - \boldsymbol{X}_2)\boldsymbol{\beta}_2$$

thus the bias in our predictions also depends on the DOE used, the model fit, and on the true model the process follows. It is this bias that the lack of fit test detects.

Example 1. Suppose the model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$. The DOE used is a 2^{3-1} and assume the true process is $Y = \beta_0 + \sum_j \beta_j x_j + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon$.

Then,

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

and

$$\boldsymbol{X}_2 = \begin{bmatrix} 1 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Thus, $\boldsymbol{X}_1'\boldsymbol{X}_1 = 4\boldsymbol{I}$, $(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1} = \frac{1}{4}\boldsymbol{I}$ and

$$\boldsymbol{X}_1'\boldsymbol{X}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 4 \\ 0 & 4 & 0 \\ 4 & 0 & 0 \end{bmatrix}.$$

Finally, the alias matrix is

$$\boldsymbol{A} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and from $E[\widehat{\boldsymbol{\beta}_1}] = \boldsymbol{\beta}_1 + \boldsymbol{A}\boldsymbol{\beta}_2$ we get that

$$E\begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \end{bmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} 0 \\ \beta_{23} \\ \beta_{13} \\ \beta_{12} \end{pmatrix}.$$

Therefore, the intercept estimate is unbiased, but the main effects are biased, their bias equal to the 2-factor interaction not involving the factor corresponding to the main effect. This is exactly what we get when we look at the defining relation of the experiment ($I = ABC$, from which $A = BC, B = AC$ and $C = AB$). ■

The previous example does not show the value of computing an Alias matrix since the DOE was a 2-level factorial and the true model only had 2-factor interactions. In this case, the algebra of defining relations provides the aliases (and biases) with much less computation. However, the alias matrix is useful to obtain biases for DOE's and models more complicated than that, as the following example shows.

Example 2. Suppose the model is first order but the true process description is given by a second order polynomial. Using a 2^k factorial, it is easy to see from the Alias matrix that the estimates of the intercept and the pure quadratic terms will be aliased, but the main effect estimates will be unbiased. ■

Both variance and bias properties of the fitted response should be considered when modeling a process with the purpose of optimizing it. Just as in basic

statistics, the Mean Square Error function of the fitted response model gives the sum of these two components:

$$\begin{aligned} \text{MSE}(\widehat{Y}(\boldsymbol{x})) &= E\left[\left(\widehat{Y}(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] \\ &= E\left[\left(\widehat{Y}(\boldsymbol{x}) - E[\widehat{Y}(\boldsymbol{x})]\right)^2\right] + \left(E[\widehat{Y}(\boldsymbol{x}) - f(\boldsymbol{x})]\right)^2 \\ &= \text{Var}(\widehat{Y}(\boldsymbol{x})) + \left[Bias(\widehat{Y}(\boldsymbol{x}))\right]^2. \end{aligned}$$

3.8 Adding Center Points to a 2^k (or 2^{k-r}) Design

It was mentioned earlier that adding center points to a 2-level orthogonal design will not result in minimum variance parameter estimates, yet it was still suggested that the center be added and replicated for these designs. In response surface work, center points in this type of designs provide a model-independent estimate of the error variance and allow to test for curvature. In this section we provide some rationale about this test, typically described and recommend in Design of Experiments books (see Appendix A for a review of the mechanics of the test).

The test is:

$$H_0 : \sum_{i=1}^{k} \beta_{jj} = 0$$

tested versus a two sided alternative. Here recall that β_{jj} is the coefficient associated with x_j^2, a pure quadratic term in a second order model (see Chapter 1). The test statistic is based on the single degree of freedom curvature sum of squares:

$$SS_{curvature} = \frac{n_f n_c (\overline{y}_f - \overline{y}_c)^2}{n_f + n_c}$$

where n_f is the number of factorial (corner) points, n_c is the number of center point replicates, and $\overline{y}_c$ and $\overline{y}_f$ are the average observed responses in factorial and center points. This one degree of freedom sum of squares is contrasted with an estimate of σ^2 (perhaps, the estimate obtained from the n_c center points) using an F statistic:

$$F_0 = \frac{SS_{curv}}{\widehat{\sigma}^2}.$$

This statistic follows an $F_{1,v}$ distribution, where v is the degrees of freedom of $\widehat{\sigma}^2$. The intuition is that if there is curvature of a parabolic nature in the inside of the "box" defined by the DOE, the observations at the center and

at the corners will differ, so we should reject the null hypothesis when their difference is significant.

The reason for an F test is that, under H_0, $E[\overline{y}_f - \overline{y}_c] = \sum_{i=1}^{k} \beta_{ii} = 0$, so

$$\frac{\overline{y}_f - \overline{y}_c - 0}{\widehat{\sigma}\sqrt{\frac{1}{n_f} + \frac{1}{n_c}}} \sim t_v$$

and squaring this quantity we get $F_0 \sim F_{1,v}$.

There still remains the question of the relation of the statistic $\overline{y}_f - \overline{y}_c$ and the term $\sum_{i=1}^{k} \beta_{ii}$. The following statement indicates the relation.

The statistic $\overline{y}_f - \overline{y}_c$ is equal to $\sum_{j=1}^{k} \widehat{\beta}_{jj}$ (where the $\widehat{\beta}_{jj}$ are the OLS estimators) if a 2^k or 2^{k-r} factorial with center points is run to fit a full second order model.

Proof. We first consider the $\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{X}$ matrices:

$$\boldsymbol{X} = \left[\begin{array}{cccccccc|ccc} \beta_0 & \beta_1 & \beta_2 & \dots & \beta_k & \beta_{12} & \dots & \beta_{k-1,k} & \beta_{11} & \dots & \beta_{kk} \\ \hline 1 & -1 & \dots & \dots & & & & & 1 & & 1 \\ 1 & 1 & & & & & & & 1 & & 1 \\ 1 & -1 & & \pm 1 & & & & & 1 & & 1 \\ 1 & 1 & & & \ddots & & & & \vdots & \ddots & \vdots \\ \vdots & \vdots & & & & & & & 1 & & \\ \hline 1 & 0 & 0 & \dots & & & & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \ddots & & & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & 0 & 0 & \dots & \dots & 0 & 0 & \dots & 0 \end{array}\right]$$

(where we show the column variables, for clarity), and

$$\boldsymbol{X}'\boldsymbol{X} = \left[\begin{array}{ccccc|ccc} n_f + n_c & 0 & 0 & \dots & 0 & n_f & \dots & n_f \\ & n_f & 0 & \dots & 0 & & & \\ & & n_f & \dots & 0 & & 0's & \\ & & & \ddots & \vdots & \vdots & \vdots & \vdots \\ \text{symmetric} & & & & n_f & 0 & \dots & 0 \\ \hline n_f & 0 & \dots & & 0 & n_f & \dots & n_f \\ \dots & & \dots & & \vdots & \dots & \ddots & \dots \\ n_f & 0 & \dots & & 0 & n_f & \dots & n_f \end{array}\right].$$

The normal equations are $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, where the right hand side is

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_{i=1}^{nf+n_c} y_i \\ \vdots \\ \hline \sum_{i=1}^{n_f} y_i \\ \vdots \end{pmatrix}$$

The first normal equation is then given by:

$$(n_f + n_c)\widehat{\beta}_0 + n_f \sum_{j=1}^{k} \widehat{\beta}_{jj} = \sum_{i=1}^{n_f+n_c} y_i. \tag{3.4}$$

The first normal equation below the partition is given by

$$n_f\widehat{\beta}_0 + n_f \sum_{j=1}^{k} \widehat{\beta}_{jj} = \sum_{i=1}^{n_f} y_i. \tag{3.5}$$

From (3.4)–(3.5) we get

$$n_c\widehat{\beta}_0 = \sum_{\text{last } n_c \text{ terms}} y_i \quad \Rightarrow \quad \widehat{\beta}_0 = \overline{y}_c. \tag{3.6}$$

Substituting (3.6) into (3.5) gives

$$\sum_{j=1}^{k} \widehat{\beta}_{jj} = \frac{\sum_{i=1}^{n_f} y_i}{n_f} - \overline{y}_c = \overline{y}_f - \overline{y}_c \ \blacksquare$$

If quadratic effects are significant, the experimental design needs to be augmented to allow the fit of a higher order model which includes such terms. This will be discussed in the Chapter 5. We finally look in this chapter at one factor at a time experimental designs from a more advance viewpoint.

3.9 Adaptive Two Level, One-factor-at-a-time Experimentation**

In a recent and interesting paper[14], Frey and Wang [56] propose to run experiments using a 2-level strategy where the experimental runs *are* varied in a one-factor-at-a-time fashion. They coined they approach *adaptive* OFAT, since the key of the approach is that the design "returns" to the best point in the factorial design if no significant improvement has been observed in the response. The search ends when all factors have been varied at least once. Figure 3.4 illustrates this idea for an experiment with 3 factors.

[14]This section contains relative more advanced material and may be skipped without loss of continuity.

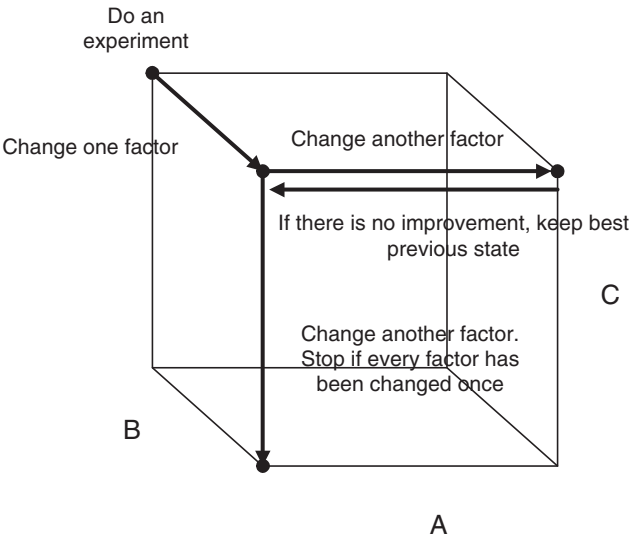


Figure 3.4. Illustration of Adaptive OFAT experimentation when $k = 3$. Adapted from Frey and Wang [56]

Frey and Wang [56] proved that there are significant improvements in the response to be gained under this strategy. Their results are based on the model they assumed:

$$y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} x_i x_j + \varepsilon \tag{3.7}$$

where

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$
$$\beta_i \sim N(0, \sigma_{ME}^2)$$
$$\beta_{ij} \sim N(0, \sigma_{INT}^2)$$

and where $x_i \in \{-1, +1\}$. In this model, the parameters are assumed to be random variables that vary from experimental *session* to experimental session. That is, once we start conducting the experiments the values of the parameters are fixed, and all the variability comes from the experimental error ε, as it is assumed in standard experimental design analysis. But before starting the experiments, the parameters are unknown (adaptive OFAT does not

assume knowledge of the "hyperparameters" $\sigma_{ME}, \sigma_{INT}$, they are introduced to analyze the procedure only). Thus, before starting any experiment, the expected response is zero, since all coefficients have mean zero and no intercept is included in the model[15]. We will comment on the appropriateness of these assumptions after we discuss some of the results found by these authors. Following Frey and Wang, we assume maximization in this section.

As one starts to conducts experiments, the model assumed in adaptive OFAT says that what we get are realizations of a stochastic process in which $E[y|\boldsymbol{\beta}]+\varepsilon$ is observed. The experimental procedure depicted in Figure 3.4 can be formalized as follows:

1 Let $O_0 = y(w_1, w_2, \ldots, w_k)$. This is the function value at the baseline point $x_i = w_i,\ i = 1, \ldots, k$.

2 Let $O_1 = y(-w_1, w_2, \ldots, w_k)$. We are varying factor 1.

3 Set $x_1 = w_1 Sgn(O_0 - O_1)$, i.e., choose the best value for x_1. Here $Sgn(z) = -1$ if $z \leq 0$ and equals $+1$ if $z > 0$.

4 Let $O_i = y(x_1, x_2, \ldots, -w_i, w_{i+1}, \ldots, w_k)$ for $i = 2, 3, \ldots, k$ (observed responses at subsequent points).

5 Set $x_i = w_i Sgn[\max(O_0, O_1, \ldots, O_{i-1}) - O_i]$ for $i = 2, 3, \ldots, k$, i.e., choose the best values for other factors by comparing with earlier observed responses.

In step 3, x_1 is simply set to whatever level provides the highest response (recall it is assumed only two levels, -1 and $+1$, are possible for each factor)[16]. Under these assumptions, Frey and Wang show some interesting results. In particular, they show that the expected improvement in the response after changing the first factor is

[15]This in inconsequential since if the intercept is included it would simply represent the expected baseline response value.

[16]This strategy is reminiscent of *greedy procedures* used in combinatorial optimization. See, e.g., [122].

$$E[y(x_1, w_2, \ldots, w_k)] = E[\beta_1 x_1] + (k-1)E[\beta_{1j} x_1 w_j]$$

where

$$E[\beta_1 x_1] = \sqrt{\frac{2}{\pi}} \frac{\sigma^2_{ME}}{\sqrt{\sigma^2_{ME} + (k-1)\sigma^2_{INT} + \sigma^2_{\varepsilon}/2}} \quad (3.8)$$

and

$$E[\beta_{1j} x_1 w_j] = \sqrt{\frac{2}{\pi}} \frac{\sigma^2_{INT}}{\sqrt{\sigma^2_{ME} + (k-1)\sigma^2_{INT} + \sigma^2_{\varepsilon}/2}} \quad (3.9)$$

Thus, if the $\sigma_{ME} \gg \sigma_{INT}$, most of the improvement can be attributed to the main effect of the first factor, as one would expect. If the opposite occurs, most of the improvement can be attributed to the two factor interactions that involve the first factor. Thus, in apparent contradiction to what is usually known about OFAT, two factor interactions are being "exploited"[17]. Based on simulations of their procedure, Frey and Wang conjecture that the average improvement after the first factor is changed is about $1/n$ of the maximum possible improvement, provided that $\sigma_{ME} > \sigma_{INT}$ and that the experimental error (σ_{ε}) is also moderate compared to σ_{ME}.

For subsequent steps, Frey and Wang show that additional improvements over (3.8–3.9) are expected from the possible exploitation of two factor interactions[18]. This apparent paradox, given that an OFAT experimental design evidently confounds two factor interactions with main effects can be explained because the *probability* of observing some improvement from this exploitation of the interactions is better than random chance (for specific probability theorems related to the exploitation of effects, see Frey and Wang's paper [56]).

The idea of adaptive optimization, in which the best result thus far is saved has not been utilized much in the experimental optimization literature. It is reminiscent of some pure search (non-model based) methods to process optimization, such as Nelder and Mead's [119] adaptive simplex search method. An interesting contribution of Frey and Wang is to point out the importance of the ability of *exploiting* an interaction, for which it is not necessary to *estimate* them "clear" of any aliases. Furthermore, being able to exploit an interaction

[17]More precisely, a two factor interaction is *exploited* if we are able to set x_i and x_j such that $\beta_{ij} x_i x_j > 0$. In this sense, adaptive OFAT starts to exploit two factor interactions only from the second step on.

[18]Frey and Wang [56] show how $P(\beta_{ij} x_i x_j > 0) \geq P(\beta_{12} x_1 x_2 > 0)$ for any i, j (with $i \neq j$). That is, the probability of exploiting interactions in later stages of the method is larger than the probability of exploiting the interaction at the second step.

is a probabilistic event. Both the expected improvement and the probability of exploiting interactions can be computed under reasonable assumptions to compare different experimental strategies for process optimization.

The main conclusion by the proponents of adaptive OFAT is that it is a technique of value if experimental noise is low. This happens to be the same situation under which Box and Wilson [29] recommended steepest ascent. Note how none of these two techniques consider two-factor interactions explicitly[19]. As mentioned by Frey and Wang, it is of interest to determine the expected improvements obtained by alternative optimization techniques. These authors compared adaptive OFAT with saturated ($n = p + 1$) resolution III factorial designs. They found that after conducting all n runs in a saturated resolution III design, the expected improvement is

$$E[y] = \sigma_{ME}^2 \sqrt{\frac{2}{\pi}} \sqrt{\sigma_{ME}^2 + \frac{\binom{k}{2}}{k} \sigma_{INT}^2 + \frac{\sigma_\varepsilon^2}{k+1}} \tag{3.10}$$

and that in this case $P(\beta_{ij} x_i x_j > 0) = 0.5$, exactly as flipping a coin ("random chance"). In contrast, the expected improvement after a complete session of adaptive OFAT is performed will be larger than (3.10) if the experimental noise is low or the interactions (σ_{INT}) large. Furthermore, the probability of exploiting interactions is better than random choice in these cases. Thus, based on these results, *if the purpose of the experiment is to optimize the response over the vertices of a hypercube*, adaptive OFAT is to be preferred to running a saturated resolution III design.

The evaluation of the response improvements rests on the assumption of the parameters being zero mean random variables before starting to conduct any experiment. This has a Bayesian interpretation (see Chapters 11 and 12). If the *a priori* variance of a main effect is very large, this is indicative of lack of knowledge about this effect before running the experiment, a situation we refer to as approaching a "noninformative" prior. Thus, the results from (3.8–3.9) indicate the intriguing result that the less we know about main effects and

[19]But also observe that steepest ascent is an "interior point" method, in the language of optimization methods, that proceeds by interpolating/extrapolating the (−1,1) factor levels, whereas adaptive OFAT is restricted to the vertices of the hypercube defined by a two level factorial.

interaction effects, the more we are to gain from them. An informative Bayesian adaptive approach in which the magnitude of the effects is bounded *a priori* – perhaps, one knows the direction of some of the effects – and/or one in which the distributions of the parameters are updated as more experiments are conducted, in a classical Bayesian way, seems as fruitful research topic based on more reasonable assumptions, Another interesting extension is to study adaptive OFAT over *sequences* of "hypercubes" in k dimensions, in a similar way that search methods operate. These extensions, however, are beyond the scope of this book.

3.10 Problems

1 Suppose we wish to fit a model for $k = 6$ with main effects and all 2-factor interactions with a 2^{6-2}_{IV} design. Compute the $\boldsymbol{X}'\boldsymbol{X}$ matrix. Is this DOE variance-optimal (do you need to invert $\boldsymbol{X}'\boldsymbol{X}$ to answer this?). What about if you use instead a 2^{6-1}_{VI} design? Explain.

2 Suppose we fit $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ using a 2^2 design (no center runs) when the true model is a full quadratic. Obtain the Alias matrix and interpret.

3 Consider the following "simplex" design in two factors:

x_1	x_2
$\sqrt{3/2}$	$-1/\sqrt{2}$
$-\sqrt{3/2}$	$-1/\sqrt{2}$
0	$2/\sqrt{2}$

a) Is this design orthogonal for a first order polynomial model?

b) What is a disadvantage of this design for fitting a first order model?

c) What are the aliases of this design if the true model has a two-factor interaction in addition to the first order model terms?.

4 Consider the following 2^{5-1} factorial experiment in five factors: ffff

x_1	x_2	x_3	x_4	x_5	y (observed)
$-$	$-$	$-$	$-$	$+$	51.8
$+$	$-$	$-$	$-$	$-$	56.3
$-$	$+$	$-$	$-$	$-$	56.8
$+$	$+$	$-$	$-$	$+$	48.3
$-$	$-$	$+$	$-$	$-$	62.3
$+$	$-$	$+$	$-$	$+$	49.8
$-$	$+$	$+$	$-$	$+$	49.0
$+$	$+$	$+$	$-$	$-$	46.0
$-$	$-$	$-$	$+$	$-$	72.6
$+$	$-$	$-$	$+$	$+$	49.5
$-$	$+$	$-$	$+$	$+$	56.8
$+$	$+$	$-$	$+$	$-$	63.1
$-$	$-$	$+$	$+$	$+$	64.6
$+$	$-$	$+$	$+$	$-$	67.8
$-$	$+$	$+$	$+$	$-$	70.3
$+$	$+$	$+$	$+$	$+$	49.8

Suppose you add four center points to the experiment with observed responses at those runs of 57, 63, 52, and 56.

a) Find the mean square for "pure error".

b) Using the following ANOVA table, test for the significance of the "second-order" terms. Are these terms important to the model? Exactly what "second order" terms are included in this analysis?

Source	d.o.f.	SS	MS
First order	5	956.90	191.38
Second order	10	150.78	15.08
Total	15	1107.69	

c) Calculate a sum of squares for "curvature". Do you think that curvature is important?

5 How well will the single degree of freedom test for curvature work for $k = 2$ if $\beta_{11} = -\beta_{22} \neq 0$? Why?

6 It was shown that $\widehat{\beta_0}$ is aliased with $\sum_{i=1}^{k} \beta_{ii}$ if we fit a first order polynomial model using a 2-level factorial but the real model is a full quadratic. If the first order model is correct, the intercept (β_0) should be the value of $\widehat{Y}$ at the design center. How can we use these facts to develop an informal (heuristic) test for curvature?

7 Give two advantages of using a first order orthogonal experimental design.

8 Consider a 2^{5-2} fractional factorial design with generators $I = ABD$ and $I = ACE$ and coded variables in the usual (-1,$+1$) scale. Suppose the fitted first degree polynomial was

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5$$

when, in fact, to obtain adequate representation over the region covered by the x's, we would need a second degree polynomial model (all pure quadratic and two-factor interactions present). What are the expected values of $b_0, b_1, \ldots, b_5$? (Note: the first generator is associated with x_4 and the second with x_5).

9 Construct a 2^9 design using Kroenecker products as discussed in Section 3.2.1. Suggestion: use Matlab.

10 Show that the relations $a * (1) = a$ and $a * a = (1)$ are satisfied for the group of treatment combinations of a 2^3 example.

11 Prove that expressions (3.8–3.9) hold for the expected improvement after setting x_1 in adaptive OFAT ([56]).

12 Prove that the expected improvement after conducting a saturated resolution III design is given by expression (3.10) ([56]).

Chapter 4

ANALYSIS AND OPTIMIZATION OF SECOND ORDER MODELS

All exact science is dominated by the idea of approximation.

—Bertrand Russell (1872–1970)

As it can be seen from previous chapters, experimental design and process optimization are two intertwined tasks. Sequences of designed experiments are frequently run to optimize a process. In traditional RSM practice, such sequences are often first order designs with center runs that allow to test for curvature. If curvature is detected, second order experimental designs and models are used as a local approximation for process optimization. In this chapter we look at optimizing a second order model. Designs used to fit these models are described in Chapter 5.

We first look at the problem of locating and characterizing the stationary point of a second order response surface in the absence of any constraints in the controllable factors. This is the so-called *canonical analysis*, a technique proposed in the Box-Wilson paper [29]. Constrained optimization schemes, in particular, Ridge Analysis, mainly developed by Hoerl [70], are presented next. The type of inferences discussed in this chapter are all subject to sampling variation in the sense that if the experiment is repeated and the model is refit, the location of the optimum will change. These statistical considerations related to second order optimization problems are discussed in Chapter 7.

It will be assumed curvature has been detected in the neighborhood of our latest set of experiments. This is the second step of an RSM study. We are

near a stationary point, and the purpose of the study at this stage is to model the process in this new region, determine the location of the stationary point, and determine its nature, whether it is a local maximum or minimum of the true response or perhaps simply a saddle point.

Experimental design strategies for fitting second order models require at least three distinct levels in each coordinate; see Chapter 5. For now, it suffices to say that the standard way to fit these models in the RSM literature has been running a so-called Central Composite Design (CCD), proposed by Box and Wilson [29] in their original paper[1]. Obvious alternatives are 3^k factorial designs, which are typically not run in practice due to the excessive number of experiments required, and 3-level fractional factorials, which have complex aliasing and are therefore less understood by practitioners than their 2-level counterparts. The CCD is composed of a 2^{k-r} fractional factorial of at least resolution V which allows us to fit all second order interactions without any aliasing of other 2-factor interactions, some center points, and $2k$ axial points that allow to fit the pure quadratic terms.

4.1 Unconstrained Optimization and Characterization of Second Order Models: "Canonical" Analysis

Suppose we have fitted the model

$$\widehat{y} = \widehat{\beta}_0 + \boldsymbol{x}'\boldsymbol{b} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$$

($\boldsymbol{x}$ in coded units) where $\boldsymbol{b}$ contains the main effects parameter estimates, and matrix $\boldsymbol{B} = \widehat{\Xi}$ is $k \times k$ and contains "pure" quadratic and two-factor interaction estimates (for notation, see Chapter 1).

As it is well-known from Calculus (see Appendix C), to find the stationary point we simply compute $\frac{\partial \widehat{y}}{\partial x_j} = 0$ for $j = 1, ..., k$ or in vector form

$$\frac{\partial \widehat{y}}{\partial \boldsymbol{x}} = \boldsymbol{b} + 2\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$$

from where

$$\boldsymbol{x}_0 = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b}. \tag{4.1}$$

[1] An example of a CCD was already shown in Chapter 1.

Substituting in the fitted model, the predicted response at the stationary point is:

$$\widehat{y}_0 = \widehat{y}(\boldsymbol{x}_0) = \widehat{\beta}_0 + \frac{1}{2}\boldsymbol{x}_0'\boldsymbol{b}.$$

Some immediate questions that arise are: what type of surface have we fitted? and what type of stationary point do we have? Answering these basic questions have become known in the RSM literature as "canonical analysis" and corresponds to the analysis of the eigenvalues and eigenvectors of matrix $\boldsymbol{B}$. This is because the Hessian of $\widehat{y}$ is $\partial^2\widehat{y}/\partial\boldsymbol{x}^2 = 2\boldsymbol{B}$.

It was noted in Chapter 1, when the "one factor at a time" method was first discussed, that it is simpler to understand a second order polynomial model if no 2-factor interactions are present. Thus, if we can transform the fitted model and avoid the 2-factor interactions in the new coordinates, easier interpretation results. The transformation consists of:

1 translation of the origin to the stationary point;

2 rotation of the axes until they coincide with the principal axes of the contours (ellipses or hyperbolas) corresponding to the fitted model.

The translation step is achieved simply by defining

$$\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{x}_0$$

thus $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{x}_0$. Substituting this into the fitted model we get

$$\begin{aligned}
\widehat{y} &= \widehat{\beta}_0 + (\boldsymbol{z} + \boldsymbol{x}_0)'\boldsymbol{b} + (\boldsymbol{z} + \boldsymbol{x}_0)'\boldsymbol{B}(\boldsymbol{z} + \boldsymbol{x}_0) \\
&= \widehat{\beta}_0 + \boldsymbol{z}'\boldsymbol{b} + \boldsymbol{x}_0'\boldsymbol{b} + \boldsymbol{z}'\boldsymbol{B}\boldsymbol{z} + \boldsymbol{x}_0'\boldsymbol{B}\boldsymbol{x}_0 + 2\boldsymbol{z}'\boldsymbol{B}\boldsymbol{x}_0 \\
&= \widehat{y}_0 + \boldsymbol{z}'\boldsymbol{b} + \boldsymbol{z}'\boldsymbol{B}\boldsymbol{z} + 2\boldsymbol{z}'\boldsymbol{B}\boldsymbol{x}_0.
\end{aligned}$$

Substituting the stationary point (equation 4.1) we get

$$\widehat{y} = \widehat{y}(\boldsymbol{z}) = \widehat{y}_0 + \boldsymbol{z}'\boldsymbol{B}\boldsymbol{z} \tag{4.2}$$

this is the model written in the translated (but not rotated) coordinates.

Since $\boldsymbol{B}$ is real and symmetric, from the spectral theorem of linear algebra there exists an orthogonal matrix $\boldsymbol{M}$ such that

$$\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & 0 \\ & & \ddots & \\ \text{symm.} & \dots & & \lambda_k \end{bmatrix} = \boldsymbol{D}$$

where the λ_i are the k real eigenvalues of $\boldsymbol{B}$ (see Appendix C). The columns of $\boldsymbol{M}$ are the eigenvectors of $\boldsymbol{B}$, which are such that $\boldsymbol{M}'\boldsymbol{M} = \boldsymbol{I}$ (i.e., the eigenvectors are orthonormal) which implies that $\boldsymbol{M}' = \boldsymbol{M}^{-1}$.

The rotation step is achieved by defining

$$\mathbb{X} = \boldsymbol{M}'\boldsymbol{z}$$

or $\boldsymbol{z} = \boldsymbol{M}\mathbb{X}$, which after substituting into (4.2) yields

$$\begin{aligned} \widehat{y} &= \widehat{y}_0 + (\boldsymbol{M}\mathbb{X})'\boldsymbol{B}(\boldsymbol{M}\mathbb{X}) \\ &= \widehat{y}_0 + \mathbb{X}'\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M}\mathbb{X} \\ &= \widehat{y}_0 + \mathbb{X}'\boldsymbol{D}\mathbb{X} \end{aligned}$$

or

$$\widehat{y} = \widehat{y}(\mathbb{X}) = \widehat{y}_0 + \sum_{j=1}^{k} \lambda_j \mathbb{X}_j^2$$

which is called the "canonical" form of the quadratic model. This is an extremely easy function to interpret:

- If all the eigenvalues $\lambda_i < 0$, then the function decreases from $\widehat{y}_0$ as one moves away from $\boldsymbol{x}_0$; we then have that $\boldsymbol{x}_0$ is a *maximum*;
- If all the eigenvalues $\lambda_i > 0$, then the function increases from $\widehat{y}_0$ as one moves away from $\boldsymbol{x}_0$; we then have that $\boldsymbol{x}_0$ is a *minimum*;
- If we have a mixture of signs in the eigenvalues, the response increases in some directions and decreases in others. Then $\boldsymbol{x}_0$ is a *saddle point*.

The canonical analysis of matrix $\boldsymbol{B}$ gives more useful information than this. Suppose $k = 2$ and $\lambda_1 \ll \lambda_2 < 0$. We will then have a maximum, but there will be *elongation* over the direction implied by the first eigenvector. The distance to a contour of $\widehat{y}$ along the rotated axis i is proportional to $1/\sqrt{|\lambda_i|}$. Thus, if one

eigenvalue is very small relative to the others, this implies that its contours will be very elongated in the direction of the i-th eigenvector. Actually, this may be good news for a process engineer, who will have the flexibility of moving the operating conditions along $\mathbb{X}_i$ while knowing that the response is expected to deteriorate little. In the limit, if an eigenvalue is exactly zero, the paraboloid degenerates into a "cylinder", called a *ridge* among RSM authors. We then have multiple optima. In higher-dimensional problems, if a pair of eigenvalues is zero, then we have a plane of optimal operating conditions, over two canonical axes, etc.

Example. Recall that in the chemical experiment of Chapter 2 the ANOVA table, obtained from using an experiment run around the coordinates $\xi_1 = 189.5$, $\xi_2 = 350$, indicated significant curvature effects. Augmenting the 2^2 factorial experiment with axial runs at $\pm\alpha = \pm\sqrt{2}$ we obtain the aforementioned central composite design (CCD), with which we can fit a second order model. The experimental results obtained at the axial points are shown in Table 4.1. From Table 4.2, the linear and (pure) quadratic effects are significant. The lack of fit tests for linear, quadratic and Cubic (third order) models fit to these data are shown in Table 4.3 and goodness of fit information is reported in Table 4.4.

The quadratic model has a larger p-value for the lack of fit test, higher adjusted R^2, and a lower PRESS statistic (see Appendix A) than the linear or cubic models, thus it seems to be the best model. The fitted quadratic equation, in coded units, is

$$\widehat{y} = 72.0 - 11.78x_1 + 0.74x_2 - 7.25x_1^2 - 7.55x_2^2 - 4.85x_1x_2.$$

A contour plot of this function (Figure 4.1) shows that it has a single optimum point in the region of the experiment.

Table 4.1. Experimental data for the axial runs, canonical analysis example

x_1	x_2	ξ_1	ξ_2	y(=yield)
-1.414	0	147.08	350	72.58
1.414	0	231.92	350	37.42
0	-1.414	189.5	279.3	54.63
0	1.414	189.5	420.7	54.18

Table 4.2. ANOVA for chemical process data (all data in central composite design used), with significance of regression tests using the extra sum of squares principle, canonical analysis example

Source	Sum of Squares	dof	Mean Square	F_0	Prob $>F_0$
Intercept	51418.2	1	51418.2		
Linear	1113.7	2	556.8	5.56	0.024
Quadratic	768.1	3	256.0	7.69	0.013
Cubic	9.9	2	5.0	0.11	0.897
Residual	223.1	5	44.6		
Total	53533.0	13			

Table 4.3. Lack of fit (LOF) tests for a linear, quadratic, and cubic models fit to the chemical experiment data, canonical analysis example

Source	Sum of Squares	dof	Mean Square	F_0	Prob $>F_0$
Linear	827.9	6	138.0	3.19	0.141
Quadratic	59.9	3	20.0	0.46	0.725
Cubic	49.9	1	49.9	1.15	0.343
Pure error	173.2	4	43.3		

Table 4.4. Goodness of fit statistics for a linear, quadratic, and a cubic model fit to the chemical experiment data, canonical analysis example

Model	Root MSE	R^2	Adj. R^2	PRESS
Linear	10.01	0.5266	0.4319	1602.02
Quadratic	5.77	0.8898	0.8111	696.25
Cubic	6.68	0.8945	0.7468	3466.71

From the fitted second order model we have that $\boldsymbol{b}' = (-11.78, 0.74)$,

$$\boldsymbol{B} = \begin{pmatrix} -7.25 & -2.425 \\ -2.425 & -7.55 \end{pmatrix}; \qquad \boldsymbol{B}^{-1} = \begin{pmatrix} -0.1545 & 0.0496 \\ 0.0496 & -0.1483 \end{pmatrix}$$

and

$$\boldsymbol{x}_0 = -\frac{1}{2} \begin{pmatrix} -0.1545 & 0.0496 \\ 0.0496 & -0.1483 \end{pmatrix} \begin{pmatrix} -11.78 \\ 0.74 \end{pmatrix} = \begin{pmatrix} -0.9285 \\ 0.3472 \end{pmatrix}$$

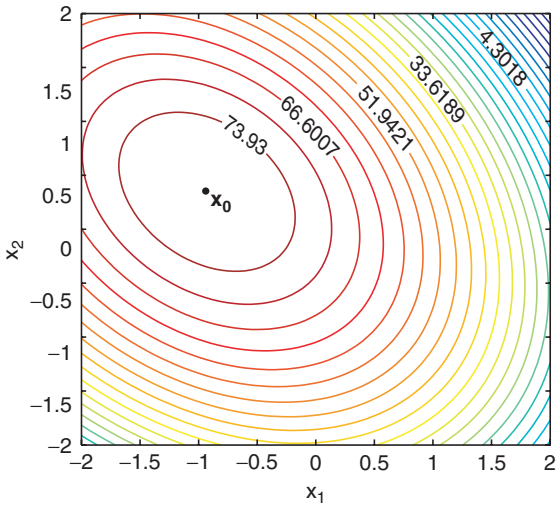


Figure 4.1. Contour plot of the fitted second order response in the Canonical Analysis example

Notice that the stationary point is well inside the experimental region. Transforming back to the original units of measurement, the coordinates of the stationary point are

$$\boldsymbol{\xi}_0 = \begin{pmatrix} 161.64\ ^\circ\text{C} \\ 367.36\ \text{min.} \end{pmatrix}.$$

The eigenvalues are the two roots of $\det(\boldsymbol{B} - \lambda \boldsymbol{I}) = (-7.25\lambda)(-7.55 - \lambda) - (-2.425(-2.245)) = 0$, which give $\lambda_1 = -4.97$ and -9.82. Thus, the stationary point is a point of maximum response. The stationary point $\boldsymbol{x}_0' = (-0.9278, 0.3468)$ corresponds to $\boldsymbol{\xi}_0' = (161.64, 367.36)$. The two normalized eigenvectors are $\boldsymbol{m}_1' = (0.7284, 0.6850)$ and $\boldsymbol{m}_2' = (-0.6850, 0.7284)$ (thus $[\boldsymbol{m}_1, \boldsymbol{m}_2] = \boldsymbol{M}$).

To get the directions of the axis of the fitted paraboloid, compute

$$\begin{aligned} \mathbb{X}_1 &= 0.7284(x_1 + 0.9278) + 0.6850(x_2 - 0.3468) \\ &= 0.43823 + 0.7284x_1 + 0.6850x_2 \end{aligned}$$

and

$$\begin{aligned} \mathbb{X}_2 &= -0.6850(x_1 + 0.9278) + 0.7284(x_2 - 0.3468) \\ &= -0.8881 + 0.6850x_1 + 0.7284x_2 \end{aligned}$$

Since $|\lambda_1| < |\lambda_2|$, there is somewhat more elongation in the $\mathbb{X}_1$ direction. This can be seen from Figure 4.1. A confirmation experiment was conducted by the

process engineer at settings $\xi_1 = 161.64$, $\xi_2 = 367.32$. The observed response was $y(\boldsymbol{\xi}_0) = 76.5$ % which is satisfactorily close to the estimated optimum. ■

4.1.1 Using SAS PROC RSREG

SAS PROC RSREG provides a nice tool for canonical analysis of second order models. The SAS commands applied to the chemical experiment data are as shown on Table 4.5. The "nocode" option was entered since the factors were already input in coded form. Note how easy it is to specify the quadratic model in the "model" line: if a list of factor names is given, the program assumes a second order polynomial in all the listed factors. This is practical if k is large. The corresponding output from SAS canonical analysis is shown in Table 4.6. As it can be seen, the results are identical to those shown earlier.

Table 4.5. SAS commands to conduct a canonical analysis

```
data;
input x1 x2 y;
cards;
−1 −1 64.33
1 −1 51.78
−1 1 77.30
1 1 45.37
0 0 62.08
0 0 79.36
0 0 75.29
0 0 73.81
0 0 69.45
−1.414 0 72.58
1.414 0 37.42
0 −1.414 54.63
0 1.414 54.18
;
proc rsreg;
model y=x1 x2 /nocode/lackfit;
run;
```

Table 4.6. SAS PROC RSREG canonical analysis output

Canonical Analysis of Response Surface		
Factor	**Critical Value**	
X1	−0.927852	
X2	0.346800	
Predicted value at stationary point 77.589146		
	Eigenvectors	
Eigenvalues	**X1**	**X2**
−4.973187	0.728460	−0.685089
−9.827317	0.685089	0.728460
Stationary point is a maximum.		

4.2 Constrained Optimization and Analysis of Second Order Models: "Ridge" Analysis

If the stationary point $\boldsymbol{x}_0$ is a maximum or a minimum but is located far away of the experimental region defined by the DOE or if it is a saddle point, it is of practical interest to locate some other operating conditions that achieve an optimum *within* the region of the experiment. In an important paper that had repercussions in the area of numerical methods, Hoerl [70] proposed to use a spherical constraint on $\boldsymbol{x}$, and called his method Ridge Analysis[2]. Assuming, as before, that a second order polynomial model is adequate in some local region, we wish to optimize

$$\widehat{y} = b_0 + \boldsymbol{b}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$$

subject to the constraint

$$\boldsymbol{x}'\boldsymbol{x} \leq \rho^2.$$

The restriction defines a spherical constraint, centered at the origin, on the operating conditions. Hoerl's approach to solve this problem is to look at the Lagrangian function and try different values of the lagrange multiplier of the constraint, μ. The lagrangian is

$$L = \widehat{y} - \mu(\boldsymbol{x}'\boldsymbol{x} - \rho^2)$$

[2]Hoerl's paper was also the first reference on the so-called *Trust Region methods*, used in mathematical programming.

thus

$$\frac{\partial L}{\partial \boldsymbol{x}} = \boldsymbol{b} + 2\boldsymbol{B}\boldsymbol{x} - 2\mu\boldsymbol{x} = \boldsymbol{0}$$

or

$$(\boldsymbol{B} - \mu\boldsymbol{I})\boldsymbol{x} = -\frac{1}{2}\boldsymbol{b} \tag{4.3}$$

where $\boldsymbol{I}$ is a $k \times k$ identity matrix. The Hessian of the Lagrangian is

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{\partial^2 L}{\partial \boldsymbol{x}^2} = 2(\boldsymbol{B} - \mu\boldsymbol{I}).$$

Therefore, depending on the value of μ we choose, we will obtain either a maximum (if μ makes $\boldsymbol{H}(\boldsymbol{x})$ negative definite) or a minimum (in case μ makes $\boldsymbol{H}(\boldsymbol{x})$ positive definite). In particular, since the eigenvalues of $(\boldsymbol{B} - \mu\boldsymbol{I})$ are equal to $\lambda_i - \mu$, $i = 1, ..., k$, we have that

- If $\mu > \max(\lambda_i)$, we obtain a *maximum* point on the sphere of radius $\rho = \sqrt{\boldsymbol{x}'\boldsymbol{x}}$.
- If $\mu < \min(\lambda_i)$, we obtain a *minimum* point on the sphere of radius $\rho = \sqrt{\boldsymbol{x}'\boldsymbol{x}}$.

In either case, the value of μ selected implies a solution $\boldsymbol{x}$ from (4.3), which in turn implies a radius ρ.

Interestingly, if $\mu = \lambda_i$, the inverse of matrix $(\boldsymbol{B} - \mu\boldsymbol{I})$ "blows up". The graph μ vs. $\sqrt{\boldsymbol{x}'\boldsymbol{x}}$ will then have "poles" at each of the eigenvalues of $\boldsymbol{B}$. Note this includes the largest and smallest eigenvalues. Thus, for values of μ greater but increasingly closer to $\max(\lambda_i)$, we will get solutions that are increasingly farther from the origin. Likewise, for values of μ smaller but increasingly closer to $\min(\lambda_i)$ will also result in increasingly farther solutions.

Example. Let us consider the example given by Hoerl [70] himself. The fitted second order polynomial is

$$\widehat{y} = 80 + 0.1x_1 + 0.2x_2 + 0.2x_1^2 + 0.1x_1^2 + x_1x_2.$$

Note from a first glance at the model that the interaction term looks relatively large. Let us assume the model fits well and proceed with the analysis. From the model we have that $\boldsymbol{b}' = (0.1, 0.2)$ and

$$\boldsymbol{B} = \begin{bmatrix} 0.2 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}.$$

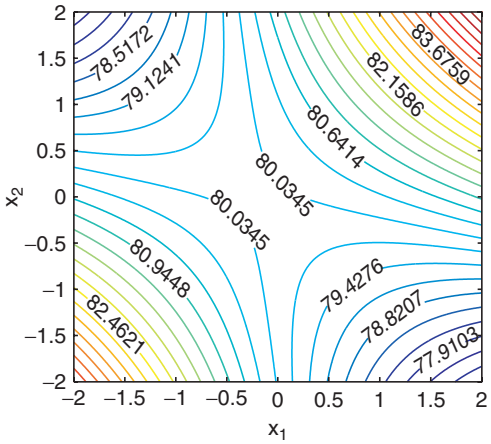


Figure 4.2. Contour plot of the fitted second order response in the Ridge Analysis example

Therefore, the stationary point is

$$\boldsymbol{x}_0 = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b} = \begin{pmatrix} -0.1917 \\ -0.0217 \end{pmatrix}.$$

The eigenvalues are $\lambda_1 = -0.3525$ and $\lambda_2 = 0.6525$, so the stationary point is a saddle point (see Figure 4.2). As mentioned earlier from our first inspection at the fitted function, the off-diagonal element of the $\boldsymbol{B}$ matrix dominates, and this results in an indefinite matrix. A ridge analysis will certainly be of benefit in this case, since saddle points, by themselves, are typically useless.

To get the point of maximum response separated a distance

$$\rho = \sqrt{\boldsymbol{x}_0'\boldsymbol{x}_0} = \sqrt{\frac{1}{4}((\boldsymbol{B} - \mu\boldsymbol{I})^{-1}\boldsymbol{b})'(\boldsymbol{B} - \mu\boldsymbol{I})^{-1}\boldsymbol{b}}$$

from the design center, we select values $\mu > \max\{\lambda_i\} = 0.6525$. To get the point of minimum response separated ρ units from the origin we set $\mu < \min\{\lambda_i\} = -0.3525$. Figure 4.3 shows a plot of ρ as a function of different values of μ.

Useful plots for a process engineer are shown in Figures 4.4–4.5. Figure 4.4 shows how the maximum and the minimum estimated response values vary over spheres of different radii ρ. Thus, this figure gives a range of average predicted values within one should expect the response to lie as we move the operating conditions away from the origin. Figure 4.5 shows the corresponding coordinates of the maximums and minimums over spheres of different radii.

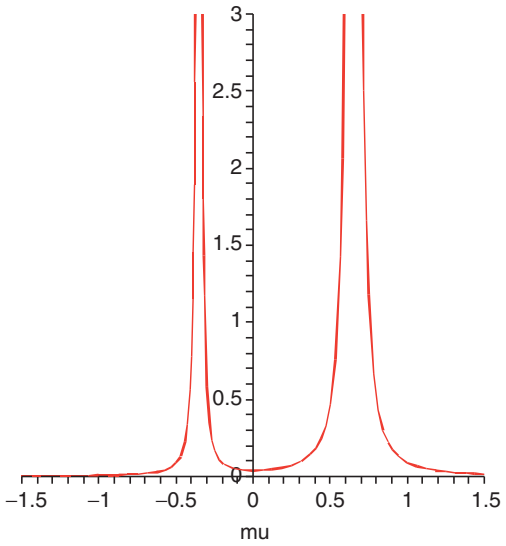


Figure 4.3. Radius (ρ) vs. values of the Lagrange multiplier μ for the example. The poles are at the two eigenvalues of $\boldsymbol{B}$

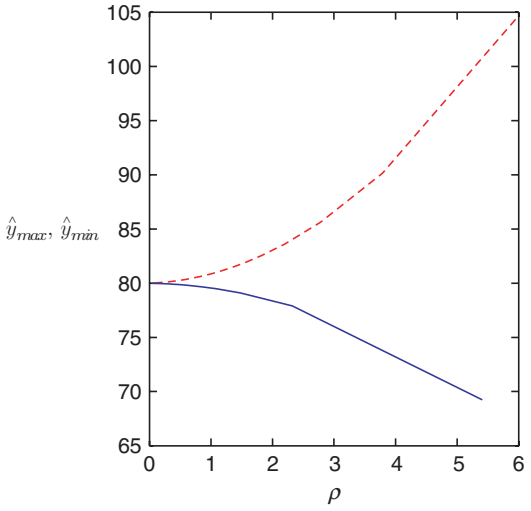


Figure 4.4. Values of the maximum and minimum estimated response on spheres of increasingly larger radius ρ

This is in agreement with Figure 4.2, since it says that to maximize y one should move approximately in the direction (1, 1) whereas to minimize one should move the process roughly in the direction $(1, -1)$. ■

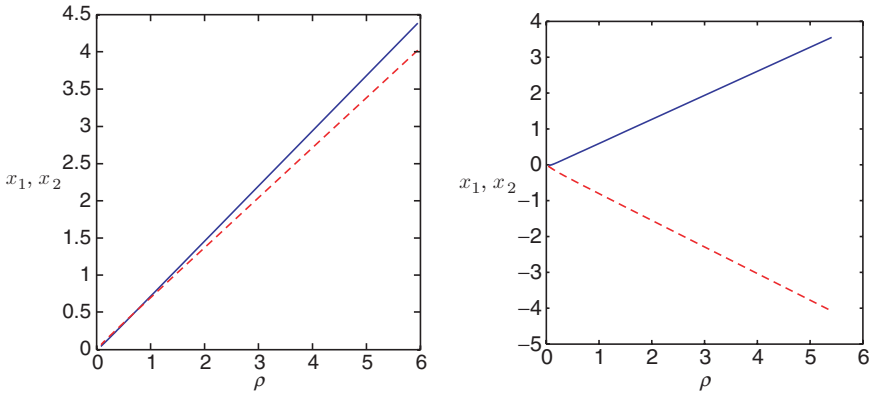


Figure 4.5. Coordinates of the maximum (left) and minimum (right) points as a function of the radius ρ

In the previous example ($k = 2$) a simple contour plot will suffice, and the rest of the analysis is not necessary. The value of the Ridge Analysis approach is that the plots in Figures 4.4–4.5 are still useful and easily obtained for $k > 2$.

As it can be seen, Canonical Analysis corresponds to unconstrained optimization and analysis of a second order polynomial, whence Ridge Analysis corresponds to constrained optimization. Box and Draper [20] (see also Draper [48]) do not show enthusiasm for Ridge Analysis.[3] They indicate that it is useless to apply a mathematical technique to data subject to sampling errors and then expect precise conclusions. This is certainly true and applies to every optimization technique in the field of RSM, including an unconstrained optimization technique such as canonical analysis. Their comment does not decrease the value of any process optimization technique provided that statistical issues are not left aside.

In Chapter 7, we discuss statistical considerations in the unconstrained and constrained optimization of response surface models.

4.3 Optimization of Multiple Response Processes

Most real life processes need to be optimized with respect to several criteria simultaneously. Frequently, operating conditions need to satisfy several conditions or constraints on m responses, $y_1, ..., y_m$. For example, in the design of a new product, product specifications need to be satisfied which determine

[3]At the time this book went to print, it appears these authors have retracted from their earlier, strong opposition to ridge analysis and now they recommend it.

the performance of the product when in use. In a manufacturing process, there are always at least two responses of interest in every process, one being quite frequently cost. This chapter discusses methods that have been developed in the Statistical literature to handle optimization problems with multiple responses.

While there has been continuous interest in academic circles to apply different multi-objective optimization techniques to solve process optimization problems as they apply in RSM, few of these have attracted the attention of Applied or Industrial Statisticians. In this chapter we focus on the most popular methods, and signal their strengths and weaknesses. We start describing the desirability method and continue with a brief description of nonlinear programming approaches, an extension of ridge analysis. While interesting, potentially useful and widely used in practice, these techniques, together with the vast majority of techniques found in academic circles for "multi-response process optimization" that simply try to translate mathematical programming techniques into a new domain miss the main point: it makes little sense to apply complicated mathematical programming techniques to models fit from data which may contain large errors. The uncertainty of the parameter estimates, and of the model for itself, needs to be accounted for. These issues will be discussed in Chapter 7 from a frequentist point of view, and later on (Part V) from a Bayesian perspective.

A method still used when handling multiple responses is simply to graph contour plots and overlay them on the space of controllable factors[4]. One then searches "by eye" a region of operating conditions which optimize all responses. Evidently this is useful, provided the number of controllable factors is 2 or maybe 3 factors. With 3 factors, interpreting the graphs is already non-trivial since one should seek on three different $x_i - x_j$ planes where the contours are projected. Evidently, for k factors, there will be a need for $\begin{pmatrix} k \\ 2 \end{pmatrix}$ such planes. Another evident limitation of this approach is that it provides the false sense that a "sweet spot" will be found with certainty, when in fact the contours are just point estimates on the mean of the responses at each location on the space of controllable factors. The contours neglect sampling variability in each response. This is an issue we will illustrate with a Bayesian approach in Chapter 12.

[4]This started as a manual graphing technique, that many statistical software packages now implement in a computer.

Other more formal approaches either try to form a new function $f : \mathbb{R}^{m \times k} \rightarrow \mathbb{R}^k$, which is then optimized. This function is supposed to contain a measure of the degrees with which each of the individual objectives or goals are being met. An instance of this approach, described in the next section, is the desirability function. Another is an approach by Khuri and Conlon [75].

A popular approach for multiple response optimization, clearly an application of mathematical programming techniques, is to regard one response as the "primary", or most important response, and treat all other $m - 1$ responses as secondary. One then optimizes the primary response subject to constraints on all secondary responses. The Dual Response approach is such a technique, which applies to the particular case of two responses. We will discuss it in Section 4.4.

4.3.1 Desirability Approach

This approach was originally proposed by Harrington [65] and later refined by Derringer and Suich [45] to its most common use in practice today.

The desirability function approach is one of the most widely used methods in industry for dealing with the optimization of multiple response processes. It is based on the idea that the "quality" of a product or process that has multiple quality characteristics, with one of them out of some "desired" limits, is completely unacceptable. The method finds operating conditions $\boldsymbol{x}$ that provide the "most desirable" response values.

For each response $y_i(\boldsymbol{x})$, a desirability function $d_i(y_i)$ assigns numbers between 0 and 1 to the possible values of y_i, with $d_i(y_i) = 0$ representing a completely undesirable value of y_i and $d_i(y_i) = 1$ representing a completely desirable or ideal response value. The individual desirabilities are then combined using the geometric mean, which gives the *overall desirability* D:

$$D = (d_1(y_1) \times d_2(y_2) \times ... \times d_m(y_m))^{1/m}$$

where m denotes the number of responses. Notice that if any response i is completely undesirable ($d_i(y_i) = 0$) then the overall desirability is zero. In practice, fitted response models $\widehat{y}_i$ are used in the method.

Depending on whether a particular response y_i is to be maximized, minimized, or assigned to a target value, different desirability functions $d_i(y_i)$ can be used. A useful class of desirability functions was proposed by Derringer and

Suich [45]. Let L_i, U_i and T_i be the lower, upper, and target values desired for response i, where $L_i \leq T_i \leq U_i$. If a response is of the "target is best" kind, then its individual desirability function is

$$d_i(\widehat{y}_i) = \begin{cases} 0 & \text{if } \widehat{y}_i(\boldsymbol{x}) < L_i \\ \left(\frac{\widehat{y}_i(\boldsymbol{x}) - L_i}{T_i - L_i}\right)^s & \text{if } L_i \leq \widehat{y}_i(\boldsymbol{x}) \leq T_i \\ \left(\frac{\widehat{y}_i(\boldsymbol{x}) - U_i}{T_i - U_i}\right)^t & \text{if } T_i \leq \widehat{y}_i(\boldsymbol{x}) \leq U_i \\ 0 & \text{if } \widehat{y}_i(\boldsymbol{x}) > U_i \end{cases}$$

where the exponents s and t determine how strictly the target value is desired. For $s = t = 1$, the desirability function increases linearly towards T_i, for $s < 1, t < 1$, the function is convex, and for $s > 1, t > 1$, the function is concave (see example below for an illustration).

If a response is to be maximized instead, the individual desirability is instead defined as

$$d_i(\widehat{y}_i) = \begin{cases} 0 & \text{if } \widehat{y}_i(\boldsymbol{x}) < L_i \\ \left(\frac{\widehat{y}_i(\boldsymbol{x}) - L_i}{T_i - L_i}\right)^s & \text{if } L_i \leq \widehat{y}_i(\boldsymbol{x}) \leq T_i \\ 1.0 & \text{if } \widehat{y}_i(\boldsymbol{x}) > T_i \end{cases}$$

where in this case T_i is interpreted as a large enough value for the response. Finally, if we want to minimize a response, we could use

$$d_i(\widehat{y}_i) = \begin{cases} 1.0 & \text{if } \widehat{y}_i(\boldsymbol{x}) < T_i \\ \left(\frac{\widehat{y}_i(\boldsymbol{x}) - U_i}{T_i - U_i}\right)^s & \text{if } T_i \leq \widehat{y}_i(\boldsymbol{x}) \leq U_i \\ 0 & \text{if } \widehat{y}_i(\boldsymbol{x}) > U_i \end{cases}$$

where T_i represents a small enough value for the response.

The desirability approach consists of the following steps:

1 Conduct experiments and fit response models for all m responses;

2 Define individual desirability functions for each response;

3 *Maximize* the overall desirability D with respect to the controllable factors.

Example. A now classical data set was used by Derringer and Suich [45] to illustrate the use of desirability optimization in an experiment arising in the development of a tire tread compound. The controllable factors are : x_1, hydrated silica level, x_2, silane coupling agent level, and x_3, sulfur level. The four responses to be optimized and their desired ranges are:

- PICO Abrasion index, y_1, $120 < y_1$;
- 200% modulus, y_2, $1000 < y_2$;
- Elongation at break, y_3, $400 < y_3 < 600$;
- Hardness, y_4, $60 < y_4 < 75$.

The first two responses are to be maximized, and the value $s = 1$ was chosen for their desirability functions. The last two responses are "target is best" with $T_3 = 500$ and $T_4 = 67.5$. The values $s = t = 1$ were chosen in both cases. The experiments were conducted according to a CCD, and the data is in Table 4.7.

Table 4.7. The Derringer-Suich data set

Run no.	x_1	x_2	x_3	y_1	y_2	y_3	y_4
1	−1.00	−1.00	1.00	102	900	470	67.5
2	1.00	−1.00	−1.00	120	860	410	65
3	−1.00	1.00	−1.00	117	800	570	77.5
4	1.00	1.00	1.00	198	2294	240	74.5
5	−1.00	−1.00	1.00	103	490	640	62.5
6	1.00	−1.00	1.00	132	1289	270	67
7	−1.00	1.00	1.00	132	1270	410	78
8	1.00	1.00	−1.00	139	1090	380	70
9	−1.63	0.00	0.00	102	770	590	76
10	1.63	0.00	0.00	154	1690	260	70
11	0.00	−1.63	0.00	96	700	520	63
12	0.00	1.63	0.00	163	1540	380	75
13	0.00	0.00	−1.63	116	2184	520	65
14	0.00	0.00	1.63	153	1784	290	71
15	0.00	0.00	0.00	133	1300	380	70
16	0.00	0.00	0.00	133	1300	380	68.5
17	0.00	0.00	0.00	140	1145	430	68
18	0.00	0.00	0.00	142	1090	430	68
19	0.00	0.00	0.00	145	1260	390	69
20	0.00	0.00	0.00	142	1344	390	70

Using ordinary least squares and standard diagnostics, the fitted responses were:

$$\widehat{y}_1 = 139.12 + 16.49x_1 + 17.88x_2 + 2.21x_3 - 4.01x_1^2 - 3.45x_2^2 - 1.57x_3^2 + 5.12x_1x_2 - 7.88x_1x_3 - 7.13x_2x_3$$

(adj. $R^2 = 0.6903$);

$$\widehat{y}_2 = 1261.13 + 268.15x_1 + 246.5x_2 - 102.6x_3 - 83.57x_1^2 - 124.82x_2^2 + 199.2x_3^2 + 69.37x_1x_2 - 104.38x_1x_3 - 94.13x_2x_3$$

(adj. $R^2 = 0.4562$);

$$\widehat{y}_3 = 417.5 - 99.67x_1 - 31.4x_2 - 27.42x_3$$

(adj $R^2 = 0.6224$);

$$\widehat{y}_4 = 68.91 - 1.41x_1 + 4.32x_2 + 0.21x_3 + 1.56x_1^2 + 0.058x_2^2 - 0.32x_3^2 - 1.62x_1x_2 + 0.25x_1x_3 - 0.12x_2x_3$$

(adj. $R^2 = 0.7466$).

Note that no interactions were significant for response 3, and that the fit for response 2 is quite poor.

Figure 4.6 displays an overlaid contour plot of the four fitted responses, prepared with the Minitab software. The unshaded region corresponds to settings for x_1 and x_2 that satisfy the response constraints.

Figure 4.7 shows the a plot of each of the responses as a function of the three controllable factors, a graph prepared with the Minitab software. The best solution found is shown. The best solution is $\boldsymbol{x}^{*'} = (-0.05, 0.14, -0.86)$ and results in $d_1(\widehat{y}_1) = 0.31 \quad (\widehat{y}_1(\boldsymbol{x}^*) = 129.4), d_2(\widehat{y}_2) = 1.0 \quad (\widehat{y}_2(\boldsymbol{x}^*) = 1299.9), d_3(\widehat{y}_3) = 0.65 \quad (\widehat{y}_3(\boldsymbol{x}^*) = 465.8)$ and $d_4(\widehat{y}_4) = 0.93 \quad (\widehat{y}_4(\boldsymbol{x}^*) = 68.02)$. The overall desirability for this solution is 0.662. All responses are predicted to be within the desired limits. ■.

Although easy to use thanks to its availability in best-selling statistical packages, the desirability function has some severe limitations. Perhaps the worst one is that it gives a sense of being able to achieve solutions that will fall inside a "sweet spot" with certainty, the same false impression that simple contours provide. The Derringer and Suich example is an excellent illustration of this point. The second response has a terrible fit, yet the authors proceeded as if a second order model fit well.

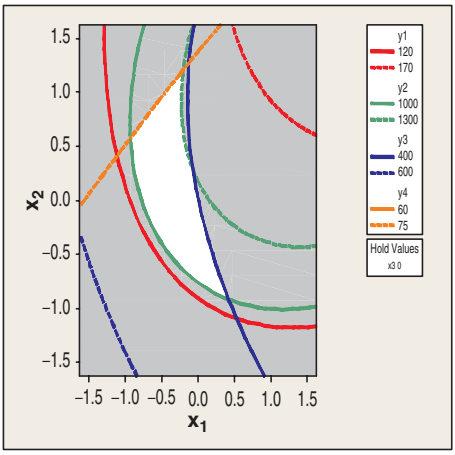


Figure 4.6. Overlaid contour plot of the four responses. Unshaded region satisfies the response constraints

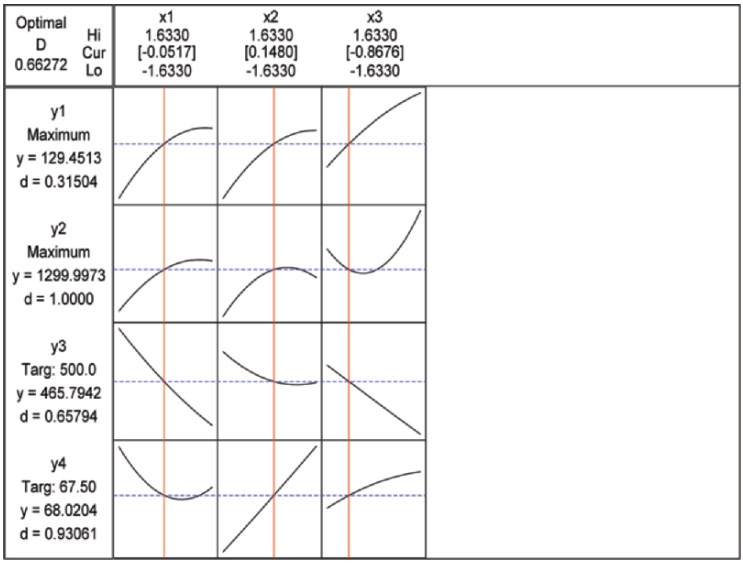


Figure 4.7. Relation between the responses and controllable factors, with optimal solution shown, Desirability example

A statistical aspect, minor compared with the seriousness of neglecting sampling variability, is that the method does not incorporate the covariances of the responses into consideration. As it will be seen in Chapter 12, a Bayesian approach is probably the best way to resolve all these difficulties.

4.4 Nonlinear Programming Approaches

The analysis of multiple response systems usually involves some type of optimization problem. When one response can be chosen as the "primary", or most important response, and bounds or targets can be defined on all other responses, a mathematical programming approach can be taken. If this is not possible, the desirability approach could be used instead.

In the mathematical programming approach the primary response is maximized or minimized, as desired, subject to appropriate constraints on all other responses. The case of two responses ("dual" responses) has been studied in more detail by some authors [42, 43, 114].

The optimization of dual response systems (DRS) consists of finding operating conditions $\boldsymbol{x}$ that

$$\begin{aligned} &\text{optimize} && \widehat{y}_p(\boldsymbol{x}) \\ &\text{subject to:} && \widehat{y}_s(\boldsymbol{x}) = T \\ & && \boldsymbol{x}'\boldsymbol{x} \leq \rho^2 \end{aligned}$$

where T is the target value for the secondary response and ρ is the radius of a spherical constraint that limits the region in the controllable factor space where the search should be undertaken. The value of ρ should be chosen with the purpose of avoiding solutions that extrapolate too far outside of the region where the experimental data were obtained. For example, if the experimental design is a central composite design, choosing $\rho = \alpha$ (axial distance) is a logical choice. Bounds of the form $L \leq x_i \leq U$ can be used instead if a cuboidal experimental region was used (e.g., when using a factorial experiment). Note that a Ridge Analysis problem is nothing but a DRS problem where the secondary constraint is absent. Thus, any algorithm or solver for DRS's will also work for the Ridge Analysis of single-response systems.

In a DRS, the response models $\widehat{y}_p$ and $\widehat{y}_s$ can be linear, quadratic or even cubic polynomials. A nonlinear programming algorithm has to be used for the optimization of a DRS. For the particular case of quadratic responses, an equality constraint for the secondary response, and a spherical region of experimentation, specialized optimization algorithms exist that guarantee global optimal solutions, regardless of the convexity or concavity of the two responses [43].

In the more general case of inequality constraints or a cuboidal region of experimentation, a general purpose nonlinear solver must be used and several starting points should be tried to avoid local optima. This is illustrated in the next example.

Example. 3 components (x_1, x_2, x_3) of a propellant need to be selected to maximize a primary reponse, burning rate (y_1) subject to satisfactory levels of two secondary reponses, namely, the variance of the burning rate (y_2) and the cost (y_3). The three components must add up to 100% of the mixture. The fitted models were:

$$\begin{aligned}
\widehat{y}_1 &= 35.4x_1 + 42.77x_2 + 70.36x_3 + 16.02x_1x_2 + 36.33x_1x_3 \\
&\quad + 136.8x_2x_3 + 854.9x_1x_2x_3 \\
\widehat{y}_2 &= 3.88x_1 + 9.03x_2 + 13.63x_3 - 0.1904x_1x_2 \\
&\quad - 16.61x_1x_3 - 27.67x_2x_3 \\
\widehat{y}_3 &= 23.13x_1 + 19.73x_2 + 14.73x_3.
\end{aligned}$$

The optimization problem is therefore:

$$\begin{array}{ll}
\text{maximize} & \widehat{y}_1(\boldsymbol{x}) \\
\text{subject to} & \widehat{y}_2(\boldsymbol{x}) \leq 4.5 \\
& \widehat{y}_3(\boldsymbol{x}) \leq 20 \\
& x_1 + x_2 + x_3 = 1.0 \\
& 0 \leq x_1 \leq 1 \\
& 0 \leq x_2 \leq 1 \\
& 0 \leq x_3 \leq 1
\end{array}$$

The solution to this problem using a nonlinear programming code is $\boldsymbol{x}^{*'} = (0.212, 0.343, 0.443)$ which provides $\widehat{y}_1 = 106.62$, $\widehat{y}_2 = 4.17$, and $\widehat{y}_3 = 18.23$. Therefore, both secondary responses meet the desired bounds. The solver should be run from a variety of starting points to avoid local optima. ■.

We should point out, once again, that all these optimization techniques are of value in case good models have been fit, hence the importance of good model building practices. These methods may have the negative effect of providing an air of authority to a solution $\boldsymbol{x}^*$ based on poorly fitted models.

4.5 Problems

1 Suppose that a central composite design was used to fit the equation:

$$\widehat{y} = 78.89 + 2.27x_1 + 3.49x_2 - 2.08x_1^2 - 2.92x_2^2 - 2.88x_1x_2$$

with adequate fit.

a) Find the coordinates of the stationary point.

b) Express the fitted equation in canonical form. Characterize the nature of the stationary point. If there is attenuation (i.e., elongation) in some direction, give the direction of the attenuation in the original variables.

c) Suppose that in b) you get $\lambda_1 \approx 0.0$ and $\lambda_2 \approx 0.0$. How can you characterize the response in the region of the stationary point?

2 Consider the following response surface in two factors:

$$\hat{y} = 80 + 0.1x_1 + 0.2x_2 + 0.2x_1^2 + 0.1x_2^2 + x_1x_2.$$

a) Find the coordinates of the stationary point.

b) What type of response function is this?

c) Perform a ridge analysis for this response. In particular, find the values of x_1 and x_2 that maximize the response on a sphere of radius approximately equal to $\sqrt{3}$.

3 Consider the Hardness response in the example in Section 4.3.1 (desirability approach).

a) Perform a canonical analysis on the hardness response. What type of response is this?

b) Find the coordinates of the stationary point

c) Perform a ridge analysis on the hardness response. Draw the radius vs. lagrange multiplier plot for this response, similar to Figure 4.3.

4 Consider the response surface in two factors:

$$\hat{y} = 80 + 0.3x_1 - 0.2x_2 + 0.5x_1^2 + 0.2x_2^2 + 1.2x_1x_2.$$

a) Find the coordinates of the stationary point.

b) What type of response function is this?

c) Perform a ridge analysis for this response. In particular, find the values of x_1 and x_2 that maximize the response on a sphere of radius approximately equal to $\sqrt{3}$.

5 Consider the Derringer and Suich example (Section 4.3.1). Solve a nonlinear programming approach where y_1 is maximized subject to $y_2 > 1000$, $400 < y_3 < 600$, and $60 < y_4 < 75$. Compare to the solution obtained using the desirability approach.

6 Consider the machining example in Section 1.3.1. In that example, the stationary point is a saddle. Therefore, find the point of maximum tool life within a spherical region of radius $\sqrt{3}$ in coded units.

7 For the machining example in Section 1.3.1, draw figures similar to Figures 4.3 to 4.5.

8 Refer to the response fitted in Problem 2. Draw figures similar to Figures 4.3 to 4.5 for this fitted model.

9 Refer to the response fitted in Problem 1. Draw figures similar to Figures 4.3 to 4.5 for this fitted model.

10 Refer to the response fitted in Problem 4. Draw figures similar to Figures 4.3 to 4.5 for this fitted model.

11 Repeat Problem 2 for the fitted response:

$$\widehat{y} = 82.17 - 1.01x_1 - 8.61x_2 + 1.40x_1^2 - 8.76x_2^2 - 7.20x_1x_2.$$

Chapter 5

EXPERIMENTAL DESIGNS FOR SECOND ORDER MODELS

A theory has to be simpler than the data it tries to explain, otherwise it does not explain anything.

—G.W. Liebniz, 1646–1716.

Response Surface Methods suggest to estimate a second order polynomial when there is evidence that the response is curved in the current region of interest, or when lack of fit tests point to an inadequacy of the a first order model. The decision for when to change from using first order designs and models to second order designs and models is therefore based on the single degree of freedom test for curvature and the lack of fit (LOF) tests explained earlier. In this chapter we provide a description of designed experiments with which we can fit the second order model

$$\widehat{y} = \beta_0 + \boldsymbol{b}\boldsymbol{x}' + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}. \tag{5.1}$$

There is a wide variety of such experimental designs, some more adequate when the region of interest is cuboidal and some when the experimental region is spherical. This chapter concentrates on some of the second order DOEs most frequently used in practice. Criteria to evaluate second order designs, such as rotatability, orthogonal blocking and prediction variance are explained. As it will be seen, rotatability is an adequate criterion to evaluate designs for which we assume a spherical region of interest. We make use of the Variance Dispersion Graph (VDG), a graphical tool used to evaluate the prediction properties

of different second order experimental designs in a succinct manner. We start by describing the rotatability criterion for second order designs used when the region of interest is spherical.

5.1 Rotatable Designs

Box and Hunter [24] considered the problem of selecting practical experimental designs for second order response surfaces[1]. A question these authors tried to answer is: if no prior knowledge exists about the orientation of the response surface, how shall the design be oriented? It is obvious that the same design may have different properties if rotated. We would like to find a design that has properties (variance properties, in this case) that remain invariant with respect to rotations. The question of which designs have invariant variance properties under rotation is answered by the class of rotatable designs.

DEFINITION 5.1 *An experimental design is said to be rotatable if Var*$(\widehat{y}(\boldsymbol{x}))$ *is constant at all points* $\boldsymbol{x}$ *equidistant from the design center.*

Recall that $\text{Var}(\widehat{y}(\boldsymbol{x}))/\sigma^2 = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}$. When plotted against $\boldsymbol{x}$, this function should result in spherical contours if the design is rotatable. The prediction variance at any of the points on the *same* sphere will be constant. Thus, if the design is rotated, the prediction variance contours remain spherical, and therefore unchanged. One has to keep in mind that, in a rotatable design, points on *different* spheres (spheres of different radii) may still yield quite different predicted variance values, so a rotatable design per se is not enough to guarantee a nice variance distribution over the experimental region. The ultimate objective of a good experimental design, from a prediction variance perspective, is that the variance of the predictions is as flat and as low as possible. This will give no preference to a region over another, so if the optimum is located in a particular region, predictions in that region will be as good as those made elsewhere. We will see more recent developments on how to capture the flatness of the $\text{Var}(\widehat{y}(\boldsymbol{x}))$ function in later sections in this chapter.

The development of the rotatability criterion is also related to the fact that, for a second order model, the $\boldsymbol{X}'\boldsymbol{X}$ matrix cannot be diagonal, i.e., orthogonal designs are impossible since not all the regressors $x_0 = 1$, x_i, x_i^2 and $x_i x_j$

[1]This paper actually introduced many of the topics in this chapter, e.g., rotatability, uniform precision, and orthogonal blocking.

are functionally independent[2]. Recall that although orthogonal designs for a first order model have nice optimality properties, they also have the practical advantage, which was much more important 50 years ago than now, of making the computation of the inverse of $\boldsymbol{X}'\boldsymbol{X}$ a trivial matter. Thus, seeking a class of second order designs that simplify the structure of the $\boldsymbol{X}'\boldsymbol{X}$ was also seen as desirable from a numerical point of view.

It appears that the original motivation behind the rotatability criterion was precisely to obtain a design with a prediction variance invariant to rotations and with a simple $\boldsymbol{X}'\boldsymbol{X}$ matrix that approaches inasmuch as possible a diagonal matrix [24]. A further reason Box and co-workers (and considerable work in RSM up to the early 1990's) concentrated in rotatability is an interesting result due to Box and Draper [22, 21] that relates to *bias*. Suppose we fit a polynomial response surface model of order d_1 but the true response is a polynomial $d_2 > d_1$. These authors show that the bias of the fitted response averaged over a *spherical* region of interest R is minimized by a design that is rotatable in addition of having other properties. This is true for any orders d_2, d_1 with $d_2 > d_1$. We will comment in some more detail this result in Chapter 8. It is important, for the time being, to emphasize that these desirable bias properties hold only for a spherical region.

Having explained the origins of the rotatability criterion and its relevance, we now proceed to describe it in more detail, mainly following the aforementioned papers and authors.

In this section, we adopt the coding convention used by Box and Hunter [24]:

$$x_{iu} = \frac{\xi_{iu} - \overline{\xi}_i}{s_i}, \quad s_i^2 = \frac{\sum_{u=1}^{N}(\xi_{iu} - \overline{\xi}_i)^2}{N}, \quad i = 1, \ldots, k$$

which results in

$$\sum_{u=1}^{N} x_{iu} = 0, \quad \sum_{u=1}^{N} x_{iu}^2 = N,$$

thus the "size" of the design in the space of the controllable factors is fixed by the coding convention. This coding convention gives the $\boldsymbol{X}'\boldsymbol{X}/N$ matrix in correlation form.

[2]For a second order design that has an intercept (thus $x_{0j} = 1$ for all j), there will always be a nonzero sum of squares in the $x_0 x_i^2$ cross terms which equals to $\sum_{u=1}^{N} x_{ui}^2$. Note this appears in an off-diagonal position within the $\boldsymbol{X}'\boldsymbol{X}$ matrix.

Whether or not a design is rotatable is determined by the design moments of order through $2d$, where d is the order of the polynomial model. The experimental design is thus seen as a distribution of points. The first four moments of the distribution indicate measures of location, dispersion, skewness and kurtosis. For a first order model, the moment matrix has moments of order 1 and 2 only. The matrix $\frac{\boldsymbol{X}'\boldsymbol{X}}{N}$ collects the moments and is thus called the **moment matrix** of a design. For a first order model ($d = 1$), the moment matrix is:

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{N} = \begin{bmatrix} 1 & [1] & [2] & \dots & [k] \\ & [11] & [12] & \dots & [1k] \\ & & [22] & \dots & [2k] \\ & & & \ddots & \\ \text{symmetric} & & & & [kk] \end{bmatrix}$$

The *bracket notation* used to denote the design moments in the above is as follows:

$$[ij] = \frac{1}{n}\sum_{u=1}^{N} x_{iu}x_{ju}$$

and

$$[i] = \sum_{u=1}^{N} x_{iu}.$$

In case the Box-Hunter coding convention is used, the moment matrix for a first order model is:

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{N} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ & 1 & [12] & \dots & [1k] \\ & & 1 & [23]\dots & [2k] \\ & & & \ddots & \\ \text{symmetric} & & & & 1 \end{bmatrix}.$$

For a second order model ($d = 2$), the moment matrix has moments of orders 1 through 4. For example, for $k = 2$ (two factors), the moment matrix for a second order model is:

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{N} = \left[\begin{array}{c|cccccc} & 1 & x_1 & x_2 & x_1^2 & x_2^2 & x_1x_2 \\ \hline 1 & & [1] & [2] & [11] & [22] & [12] \\ & [11] & [12] & [111] & [122] & [112] \\ & & [22] & [112] & [222] & [122] \\ & & & [1111] & [1122] & [1222] \\ & & & & [2222] & [1222] \\ \text{symmetric} & & & & & [1122] \end{array}\right]$$

where each column has been labelled for clarity. Here, the fourth order moments are of the form

$$[1111] = \frac{1}{N}\sum_{u=1}^{N} x_{1u}^4 \quad \text{(pure fourth order moment)}$$

$$[1112] = \frac{1}{N}\sum_{u=1}^{N} x_{1u}^3 x_{2u} \quad \text{(mixed fourth order moment)},$$

etc. In general, when using an experimental design for fitting a response model of order d in k factors, a moment of order δ is given by:

$$\frac{1}{N}\sum_{u=1}^{N} x_{1u}^{\delta_1} x_{2u}^{\delta_2} \ldots x_{ku}^{\delta_k} \doteq [1^{\delta_1} 2^{\delta_2} \ldots k^{\delta_k}]$$

where $\delta = \sum_{i=1}^{k} \delta_i$.

Box and Hunter [24] showed that a necessary and sufficient condition for an experimental design to be rotatable for a polynomial model of order d is that

$$\text{design moments} = \text{moments of a spherical normal distribution}$$

for all moments up to order $2d$. From this, necessary and sufficient conditions for rotatability of a second order model ($d = 2$) were derived and are:

$$[1^{\delta_1} 2^{\delta_2} \ldots k^{\delta_k}] = \begin{cases} \dfrac{\lambda_\delta \prod_{i=1}^{k} \delta_i!}{2^{\delta/2} \prod_{i=1}^{k} \left(\frac{\delta_i}{2}\right)!} & \text{if all } \delta_i \text{ are even} \\ 0 & \text{if any } \delta_i \text{ is odd} \end{cases} \tag{5.2}$$

The quantity λ_δ depends only on δ and is a parameter that can be chosen by the experimenter. Note how, e.g., $[1112] = [1^3 2]$ implies $\delta_1 = 3$,$\delta_2 = 1$, so this is an odd moment; likewise, $[22]$ is an even moment, etc.

Before looking in more detail at designs for second order models, let us first consider how the rotatability conditions work for a first ($d = 1$) order model. Suppose we have $k = 2$ factors. Then the moment matrix reduces to:

$$\frac{\boldsymbol{X}'\boldsymbol{X}}{N} = \begin{bmatrix} 1 & [1] & [2] \\ & [11] & [12] \\ symm. & & [22] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ symm. & & 1 \end{bmatrix}$$

where the zeroes on the first row come from the coding convention or from the rotatability condition, $[12] = 0$ is from the rotatability condition (odd moment), and the ones in the diagonal from the coding convention. This seems to imply that a rotatable design is orthogonal for a first order model, that is rotatability $\Rightarrow$ orthogonality for a first order model. In fact, the rotatability conditions indicate that $[ii] = \lambda_2$, where λ_2 is an arbitrary constant. If $\lambda_2 = 1$ we are setting the diagonal elements of $\boldsymbol{X}'\boldsymbol{X}$ equal to N, and this is the same as imposing a constraint of the form $\sum_{u=1}^{N} x_{iu}^2 = c_i = N$, which can be met in various ways (see Chapter 3). Thus, for a first order model, a rotatable design will be one that is variance optimal according to Theorem 3.1.

Showing the inverse relation, that for a first order model variance optimality implies rotatability, is left as an exercise.

For a second order model ($d = 2$), the rotatability conditions (5.2) imply the following in the fourth order moments:

$$[iiii] = \frac{\lambda_4 4!0!}{2^2 2!0!} = 3\lambda_4$$

and

$$[iijj] = \frac{\lambda_4 2!2!}{2^2 1!1!} = \lambda_4$$

thus we have that the rotatability conditions for a second order model are:

a) $[iiii] = 3[iijj]$ for all i, j;

b) all odd moments through order 4 must be zero.

The rotatability condition does not depend on the scaling of the design, i.e., it does not depend on the second order moments of the design. With these

conditions, the moment matrix of a rotatable design used to fit a second order model is:

1	x_1	$\ldots$	x_k	x_1^2	x_2^2	$\vdots$	x_k^2	x_1x_2	x_1x_3	$\ldots$	$x_{k-1}x_k$
1	0	$\ldots$	0	1	1	$\ldots$	1	0	0	$\ldots$	0
	1	$\ldots$	0	0	0	$\ldots$	0	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$\ddots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$				
			1	0	0		0	0	0	$\ldots$	0
				$3\lambda_4$	λ_4	$\ldots$	λ_4	0	0	$\ldots$	0
					$3\lambda_4$	$\ddots$	λ_4	$\vdots$	$\vdots$	$\vdots$	$\vdots$
						$\ddots$	λ_4				
							$3\lambda_4$	0	0	$\ldots$	0
								λ_4	0	$\vdots$	0
									λ_4	$\ldots$	0
										$\ddots$	0
symm.											λ_4

With this structure, the inverse of the moment matrix, $N(\boldsymbol{X}'\boldsymbol{X})^{-1}$, is easy to get[3] as in Table 5.1 ([24]).

The λ_4 parameter on the table can be chosen based on a criterion other than rotatability. For example, Box and Hunter [24] showed how in a rotatable second order design the *bias* of the $\widehat{\beta}_i$ parameter estimates, when the true model is a third order polynomial, is proportional to λ_4. From (5.1), it can be shown that

$$\text{Var}(\widehat{y}(\rho)) = A[2(k+2)\lambda_4^2 + 2\lambda_4(\lambda_4-1)(k+2)\rho^2 + ((k+1)\lambda_4 - (k-1))\rho^4]$$

where $\rho = \sqrt{\boldsymbol{x}'\boldsymbol{x}}$. Furthermore,

$$\begin{aligned} \text{Cov}(b_{ii}, b_{jj}) &= (1-\lambda_4)A\sigma^2/N, \quad i \neq j; \\ \text{Cov}(b_0, b_{ii}) &= -2\lambda_4 A\sigma^2/N, \quad i = 1, \ldots k \end{aligned}$$

(all other covariances are zero). These equations imply that the choice:

$$\lambda_4 = \frac{k}{k+2}$$

[3]Note that although $\sum_{u=1}^{N} x_{ou}x_{iu} = \sum_{u=1}^{N} x_{iu}^2 = N$, when taking the inverse, the corresponding entries in the inverse of the moment matrix, shown here, actually differ.

Table 5.1. The $N\ (\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix for a Central Composite Design, with $A = (2\lambda_4[(k+2)\lambda_4 - k])^{-1}$ [22]

1	x_1	$\ldots$	x_k	x_1^2	$\ldots$	x_k^2	x_1x_2	$\ldots$	$x_{k-1}x_k$
$2\lambda_4^2(k+2)A$	0	$\ldots$	0	$-2\lambda_4 A$	$\ldots$	$-2\lambda_4 A$	0	$\ldots$	0
	1	0	0	0	0	0	0	0	0
		$\ddots$	0	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
			1	0	0	0			
				$[(k+1)\lambda_4 - (k-1)]A$	$(1-\lambda_4)A$	$(1-\lambda_4)A$			
					$\ddots$	$\vdots$			
	symmetric					$[(k+1)\lambda_4 - (k-1)]A$	0	0	0
							$1/\lambda_4$	$\ldots$	0
								$\ddots$	0
									$1/\lambda_4$

causes $A \to \infty$ thus Var$(\widehat{y})$ diverges ($\boldsymbol{X}'\boldsymbol{X}$ is not invertible in this case). This occurs if all points in the design are equidistant from the origin. Another λ_4 choice of interest is

$$\lambda_{4\,\text{ORTH}} = 1$$

which causes $\text{Cov}(b_{ii}, b_{jj}) = 0$, thus the resulting design is "orthogonal" (the $\boldsymbol{X}'\boldsymbol{X}$ matrix is not completely diagonal, since the intercept and the quadratic effects will still be correlated). As λ_4 increases over 1.0, the bias in the main effects increases.

It was mentioned before that what really matters is that the variance function be as low and flat as possible everywhere in the experimental region. Noticing this, Box and Hunter proposed to choose the value of λ_4 such that the prediction variance at the origin equals the prediction variance at a distance of 1.0 coded units, that is, choose λ_4 such that

$$\text{Var}(\widehat{y}|\rho = 0) = \text{Var}(\widehat{y}|\rho = 1).$$

These authors called such designs **uniform precision** rotatable designs. It can be shown that the value of λ_4 that satisfies this condition is

$$\lambda_{4\,(\text{UP})} = \frac{k + 3 + \sqrt{9k^2 + 14k - 7}}{4(k+2)} \tag{5.3}$$

By solving this expression for various values of k we obtain Table 5.2. Note that the values of λ_4 obtained from this formula and on the table are only valid if the Box-Hunter coding convention is used.

Note that all the $\lambda_{4\,(\text{UP})}$ values are between $k/(k+2)$ and 1.0.

We now describe how the rotatability conditions are achieved in central composite designs.

Table 5.2. Values of the $\lambda_{4\,(\text{UP})}$ parameter providing uniform precision

k	1	2	3	4	5	6	7	8	9
λ_4	0.6666	0.7844	0.8385	0.8705	0.8918	0.9070	0.9185	0.9274	0.9346
k	10	11	12	13	14	15	16	17	18
λ_4	0.94042	0.9453	0.9495	0.9530	0.9561	0.9588	0.9612	0.9634	0.96528

5.2 Central Composite Designs

Designs that can fit model (5.1) must have at least three different levels in each variable. This is satisfied by Central Composite Designs (CCD)'s, which have five levels per variable. First proposed by Box and Wilson [29] (see also [24]), these designs are formed or *composed* by three parts:

1 2^k or 2^{k-r}_R ($R \geq V$) factorial experiment (we will refer to this as the factorial part);

2 $2k$ axial points;

3 n_0 replicates of the center point.

A resolution of at least V allows clear estimates of all main effects and 2-factor interactions. The axial points allow the estimation of the pure quadratic effects. Some center points can be designed to be run together with the factorial points and some more can be run together with the axial points.

Table 5.3 shows a central composite design table. The axial distance is α. The total number of runs N is thus $F + n_0 + 2k$, where F is the number of

Table 5.3. The $\boldsymbol{D}$ matrix for a central composite design

x_1	x_2	x_3	...	x_k
± 1 from a 2^k or 2^{k-r}_V design				
0	0	0	0	0
0	0	0	0	0
⋮	⋮	⋮	⋮	⋮
0	0	0	0	0
$-\alpha$	0	0	...	0
α	0	0	...	0
0	$-\alpha$	0	...	0
0	α	0	...	0
0	0	$-\alpha$	0	0
0	0	α	0	0
⋮	⋮	⋮	⋱	⋮
0	0	0	0	$-\alpha$
0	0	0	0	α

factorial points. Notice the structure of the experimental design is fixed; the tuning parameters of this experimental design are α and n_0 (and the way the center points are allocated, as will be seen below). These can be selected to achieve particular properties. Box and Hunter suggested to choose α to induce rotatability and choose n_0 to try to get uniform precision. If the experiment is run in blocks, the allocation of center points to the different blocks of the design can be used to make the block effects orthogonal to the effects in the second order model.

If we look only at the pure quadratic columns of matrix $\boldsymbol{X}$ for a CCD, we get

x_1^2	x_2^2	$\ldots$	x_k^2
1	1	$\ldots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	$\ldots$	1
0	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	0	$\ldots$	0
α^2	0	$\ldots$	0
α^2	0	$\ldots$	0
0	α^2	$\ldots$	0
0	α^2	$\ldots$	0
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
0	0	0	α^2
0	0	0	α^2

It is easy to check that, for all $i = 1, 2, \ldots, k$:

$$\sum_{u=1}^{N} x_{iu}^2 = F + 2\alpha^2$$

$$\sum_{u=1}^{N} x_{iu}^4 = F + 2\alpha^4$$

and

$$\sum_{u=1}^{N} x_{iu}^2 x_{ju}^2 = F.$$

Recall that the rotatability condition is that $[iiii] = 3[iijj]$ and all odd moments in the moment matrix must equal to zero. The second condition, on the odd moments, is satisfied by all CCD's in the usual coding convention. Thus, we seek a value of α such that the first condition is also true. We have that

$$[iiii] = \frac{1}{N}(F + 2\alpha^4)$$

and

$$[iijj] = \frac{F}{N}$$

Thus $\frac{1}{N}(F + 2\alpha^4) = 3F/N$, from which the rotatable choice of the axial distance in a CCD is

$$\alpha = F^{1/4}.$$

It could be thought that the rotatable choice will place the design points on a sphere. This is true only when $\alpha = \sqrt{k}$. Recall that the rotatability condition matches design moments with those of the spherical distribution only up to order $2d$. The $\alpha = \sqrt{k}$ choice is rotatable only for $k = 2$ and $k = 4$.

A popular choice of the axial distance is to use $\alpha = 1$, giving what is called a *Face-centered* central composite design. This is a design with 3 levels only, an advantage from a practical point of view.

In a CCD, we vary the number of center points, n_0, to achieve uniform precision. Since the values of the λ_4 parameter that give uniform precision (given in Table 5.3) depend on Box and Hunter coding convention, we need to find the relation of these values to the design written in the usual (−1,1) coding convention. For a CCD, the usual coding satisfies $\sum_{u=1}^{N} x_{iu} = 0$ but does not satisfy $\sum_{u=1}^{N} x_{iu}^2 = N$. Therefore, consider satisfying this constraint by introducing a scaling factor g that would transform the factors from the (−1,1) convention to the BH coding convention:

$$\sum_{u=1}^{N} (gx_{iu})^2 = N$$

where the x_{iu}'s are the factors in the usual coding. Thus,

$$g^2 \sum_{u=1}^{N} x_{iu}^2 = g^2(F + 2\alpha^2) = N$$

which, solving from g gives the scaling factor necessary to meet the sum of squares constraint:

$$g = \sqrt{\frac{N}{F + 2\alpha^2}}.$$

With this, and the rotatability choice $\alpha = F^{1/4}$, the pure fourth order moment is

$$[iiii] = \frac{g^4(F + 2\alpha^4)}{N} = \frac{3N}{F + 4F^{1/2} + 4}.$$

From the rotatability condition,

$$3[iijj] = 3\lambda_4 = [iiii] \quad \Rightarrow \quad \lambda_4 = \frac{N}{F + 4F^{1/2} + 4}$$

where $N = F + 2k + n_0$. Thus, using this equation for λ_4, we modify n_0 to get as close as possible to the value λ_4 on Table 5.3, obtained from (5.3). Thus we wish to find an integer value of n_0 that solves

$$\frac{F + 2k + n_0}{F + 4F^{1/2} + 4} = \frac{k + 3 + \sqrt{9k^2 + 14k - 4}}{4(k + 2)}.$$

Integer solutions of this equation will be the exception, but we can still use an integer solution as close as possible to the value that solves it exactly. This will provide rotatable, *near* uniform precision designs.

The variance properties of CCD (and other second order) designs will be studied using variance dispersion graphs in Section 5.6. Here we want to point out that not all terms in the second order model will be estimated with the same precision or efficiency. We define the efficiency with which parameter β is estimated as

$$\frac{\sigma^2/N}{\text{Var}(\widehat{\beta})} = \frac{1}{N\text{Var}(\widehat{\beta})/\sigma^2}$$

The efficiencies are therefore measuring the variance of the parameter estimate compared to the variance we will get from the average of the N observations at a single point. The larger this number is, the more efficient the parameter is being estimated, where efficiencies can very well be greater than one. Table 5.4 shows the efficiencies for each of the terms relevant for optimization (the intercept is not), assuming that we either use the orthogonal choice of λ_4 or the uniform precision choice.

Table 5.4. Efficiencies of the second order parameter estimates obtained with a rotatable CCD for both an orthogonal or a uniform precision choice of λ_4

k	1st order	2-factor interactions		Pure quadratic	
		Orthogonal	Uniform Precison	Orthogonal	Uniform Precision
2	1	1	0.7844	2	1.3187
3	1	1	0.8385	2	1.4770
4	1	1	0.8705	2	1.5744
5	1	1	0.8918	2	1.6408
6	1	1	0.9070	2	1.6891
7	1	1	0.9185	2	1.7258
8	1	1	0.9274	2	1.7548
9	1	1	0.9346	2	1.7782
10	1	1	0.9404	2	1.7975
11	1	1	0.9453	2	1.8137
12	1	1	0.9495	2	1.8275
13	1	1	0.9530	2	1.8394
14	1	1	0.9561	2	1.8498
15	1	1	0.9588	2	1.8589
16	1	1	0.9612	2	1.8669
17	1	1	0.9634	2	1.8741
18	1	1	0.9653	2	1.8806
19	1	1	0.9670	2	1.8864
20	1	1	0.9686	2	1.8917

We note from Table 5.4 that the rotatable CCD estimates the pure quadratic terms very efficiently compared with two factor interactions and main effects.

5.3 Blocking in Second Order Models

If the factorial and axial runs of a CCD are performed in chronological order, then some block effect due to time might be present. This is common in RSM, in which the factorial part of the CCD is run first as part of the end of a steepest ascent/descent search, and, once curvature is detected, the axial runs are added. The presence of a block effect may bias the parameter estimates of the quadratic function we wish to fit. A design in which the ordinary least squares estimators of the β's are independent of the block effects δ_l is said to **block orthogonally**. Thus, the goal is to find conditions on the design that achieve orthogonal blocking.

The second order model with blocking is

$$\begin{aligned} y_u &= \beta_0 + \sum_{i=1}^{k} \beta_i x_{iu} + \sum_{i=1}^{k} \beta_{ii} x_{iu}^2 + \sum_{i<j} \sum_{j} \beta_{ij} x_{iu} x_{ju} \\ &\quad + \sum_{l=1}^{b} \delta_l z_{lu} + \varepsilon_u, \quad u = 1, 2, \ldots, n \end{aligned}$$

where b is the number of blocks (in general, there might be any number of blocks, not only two). The variable z_{lu} is a dummy or indicator variable equal to 1 if run u is performed in block l and zero otherwise.

Box and Hunter [24] derived two conditions for orthogonal blocking for a second order design that satisfies the conditions $[i] = 0$ and $[ij] = 0$. These conditions are satisfied, for example, by all rotatable designs.

Centering the dummy variables and writing $\beta_0' = \beta_0 + \sum_{l=1}^{b} \delta_l \overline{z}_l$:

$$\begin{aligned} y_u &= \beta_0' + \sum_{i=1}^{k} \beta_i x_{iu} + \sum_{i=1}^{k} \beta_{ii} x_{iu}^2 + \sum_{i<j} \sum_{j} \beta_{ij} x_{iu} x_{ju} \\ &\quad + \sum_{l=1}^{b} \delta_l (z_{lu} - \overline{z}_l) + \varepsilon_u, \quad u = 1, 2, \ldots, n. \end{aligned}$$

To have orthogonally blocking, we must first have that:

$$\begin{aligned} \sum_{u=1}^{N} x_{ui}(z_{ul} - \overline{z}_l) &= 0 \\ \Leftrightarrow \sum_{u=1}^{N} x_{ui} z_{ul} &= \overline{z}_l \sum_{u=1}^{N} x_{ui}, \quad i = 1, 2, \ldots, k, \; l = 1, \ldots, b \end{aligned}$$

Assuming $[i] = 0$, we conclude that since the right hand side must equal zero, so does the left hand side which can be written in the more compact form:

$$\sum_{u \in b_l} x_{ui} = 0, \quad i = 1, 2, \ldots, k; \; l = 1, \ldots, b \tag{5.4}$$

where b_l is the set of run indices that belong to block l. Similarly,

$$\begin{aligned} \sum_{u=1}^{N} x_{ui} x_{uj}(z_{ul} - \overline{z}_l) &= 0 \\ \Leftrightarrow \sum_{u=1}^{N} x_{ui} x_{uj} z_{ul} &= \overline{z}_l \sum_{u=1}^{N} x_{ui} x_{uj} \;\; i, j = 1, 2, \ldots, k, i \neq j, l = 1, \ldots, b \end{aligned}$$

and from assuming $[ij] = 0$:

$$\sum_{u \in b_l} x_{ui} x_{uj} = 0, \quad i, j = 1, 2, \ldots, k, i \neq j, l = 1, \ldots, b \tag{5.5}$$

Finally, in a similar way, the quadratic terms must be orthogonal:

$$\sum_{u=1}^{N} x_{ui}^2 z_{ul} = \overline{z}_l \sum_{u=1}^{N} x_{ui}^2, \quad i, j = 1, 2, \ldots, k, i \neq j, l = 1, \ldots, b$$

from which

$$\frac{\sum_{u \in b_l} x_{iu}^2}{\sum_{u=1}^{N} x_{iu}^2} = \frac{n_l}{N}, \quad i = 1, 2, \ldots, k;\ l = 1, \ldots, b \tag{5.6}$$

Summarizing: for block effects to be orthogonal, in a second order design for which $[i] = [ij] = 0$ we must have that:

1 each block must be a first-order orthogonal design as indicated by (5.4) and (5.5): (this condition is always true for CCD's in the usual two blocks);

2 the sum of squares of each variable as a fraction of the total sum of squares should equal the fraction of the number of observations with respect to the total N (as (5.6) says).

If the experimental design does not block orthogonally, the blocking effect will be confounded.

For a CCD in two blocks, factorial and axial, we have that $N = F + 2k + n_{0f} + n_{0a}$ with n_{0f} the center points allocated to the factorial part and n_{0a} the number of center points allocated to the axial block ($n_0 = n_{0f} + n_{0a}$). The second orthogonal blocking condition implies that for a CCD,

$$\frac{2\alpha^2}{F + 2\alpha^2} = \frac{2k + n_{0a}}{F + 2k + n_{0f} + n_{0a}}$$

from which, solving for α,

$$\alpha = \sqrt{\frac{F(2k + n_{0a})}{2(F + n_{0f})}} \tag{5.7}$$

which guarantees orthogonal blocking. Since this value of α depends on the allocation of center points to the two blocks, we can try to get a rotatable,

orthogonal blocking CCD by equating the orthogonal blocking choice of α to the rotatable choice of α, $F^{1/4}$, obtaining the equation

$$\frac{F^{1/2}}{2} = \frac{F + n_{0f}}{2k + n_{0a}}. \tag{5.8}$$

This equation will not always have a solution for integer n_{0f} and n_{0a}. If it does, then the value of α computed from (5.7) will yield a rotatable, orthogonally blocking CCD.

In some cases, several integer values of n_{0f} and n_{0a} will satisfy (5.8). If this happens, we then have the flexibility of making $n_{0f} + n_{0a} = n_0$ equal to the total number of center points which will give a (near) uniform precision design. This will be a highly desirable CCD; it would be rotatable, would block orthogonally, and would be a near uniform precision design.

5.4 Box-Behnken Designs

A class of 3-level designs for fitting quadratic polynomials was proposed by Box and Behnken (BB, [19]). Besides being 3-level experiments, which in itself is of practical advantage in industrial experiments compared to (say), the 5-level CCD's, these designs have other nice properties. Three useful properties are:

- these designs are either rotatable or close to rotatable, thus they are useful if the region of interest is spherical (although the experimental region is not really spherical);
- they can be run in blocks that are orthogonal to all other effects by addition of center runs;
- they change few factors at a time, which from the point of view of actually *conducting* certain experiments may be desirable if many factors are difficult to vary.

As it is discussed below, unless k is quite small the latter property actually can turn into a *disadvantage*, because it implies a relative poor estimation of the two factor interactions, as we show below. This has not been emphasized enough in the literature.

Construction of these designs is quite interesting. They are based in combining a two-level factorial with the arrangement of treatments in a *balanced incomplete block design*, or BIBD or in a *partial* BIB design, or PBIB design.

A BIBD is a single factor experimental design in which every two levels (or treatments) of the factor are tried together an equal number of runs (thus, it is balanced). However, not all levels (treatments) are varied together at each run, thus the design is incomplete. To better understand what incomplete means, it is convenient to think, as BB did, on the treatments of the factors actually as different factors. Then the design will not vary all factors simultaneously, but only certain number of factors, say $l < k$ are varied, and we leave all other $k - l$ factors constant in every run. Thus the resulting design will be closer in this sense, to a one-factor-at-a-time design (for which $l = 1$). Since not all treatments (levels) are varied within each run, this may result in a block effect. Hence the BIBD name. A PBIB design is a DOE in which not all pairs of treatments are tried the same number of times, thus the balance is partial. BB used PBIB designs with *two class associates*, meaning that each pair of treatments appears together either some m_1 or some m_2 times, where m_1 and m_2 may vary from design to design, but are constant for a given design. Same as in a BIBD, there are only $l < k$ levels varied together, so the design is incomplete and block effects may occur.

Box and Behnken proposed to combine a BIBD or a PBIB design with a 2-level factorial design in such a way that whenever there is a treatment in a block being applied, we would substitute that entry in the BIBD or PBIB design by a *whole column* of a 2^l factorial design[4], where l is the number of treatments varied in each block. The remaining $k - l$ entries that have a zero value in the BIBD or PBIB design are substituted by $2^l \times 1$ columns of zeroes.

Example. Construction of a 4 factor BB design. BB used in this case a BIBD. A BIBD design for four treatments, with two treatments varied in each block is:

	Treatments			
Block	1	2	3	4
1	*	*		
2			*	*
3	*			*
4		*	*	
5		*		*
6	*		*	

[4]In some cases, a 2^{l-r} design was used.

Substituting the columns of a 2^2 design and 4×1 column vector of zeros in the empty spaces, we get the BB design, which, following Box and Behnken can be written as:

$$\left[\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \hline \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} \\ \pm\mathbf{1} & \mathbf{0} & \mathbf{0} & \pm\mathbf{1} \\ \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} \\ \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} & \mathbf{0} \end{array}\right] \begin{array}{l} \\ \leftarrow 2^2 \\ \leftarrow 2^2 \\ \leftarrow 2^2 \\ \leftarrow 2^2 \\ \leftarrow 2^2 \\ \leftarrow 2^2 \end{array}.$$

Here, the entries $\pm\mathbf{1}$ represent the two columns of a 2^2 design. Thus this is a design in which the basic DOE is a BIBD with the columns of a 2^2 substituted 6 times. This particular design can be run in 3 blocks if to each set of 8 runs we add one center point. If we do that, the blocks will be orthogonal to all other effects of interest. The design thus has 27 runs to estimate $p = (k+1)(k+2)/2 = 15$ parameters. In addition, it is rotatable, since $[iiii] = 12/24 = 3[iijj] = 3(4)/24$. ■.

The BB designs for $k = 3$ and $k = 5$ are constructed similarly, where $l = 2$ factors are changed at a time and hence the columns of a 2^2 factorial are substituted in a BIBD[5]. For $k = 6$, the most economical BB design is based on the PBIB design[6]:

$$\left[\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \hline \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} \\ \pm\mathbf{1} & \mathbf{0} & \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \pm\mathbf{1} & \mathbf{0} & \mathbf{0} & \pm\mathbf{1} & \pm\mathbf{1} \\ \pm\mathbf{1} & \mathbf{0} & \pm\mathbf{1} & \mathbf{0} & \mathbf{0} & \pm\mathbf{1} \end{array}\right] \begin{array}{l} \\ \leftarrow 2^3 \\ \leftarrow 2^3 \\ \leftarrow 2^3 \\ \leftarrow 2^3 \\ \leftarrow 2^3 \\ \leftarrow 2^3 \end{array}$$

It is easy to see how in this design factors pairs x_1 and x_4, x_2 and x_5 and x_3 and x_6 are varied together in two blocks whereas all other pairs are varied only in one block. Hence this is a PBIB design with two associates, $m_1 = 1$ and $m_2 = 2$.

[5]This substitution is done 3 and 10 times, respectively, for $k = 3$ and $k = 5$.

[6]In [19], a much larger experiment ($N = 80$+center runs) was also suggested based on a BIBD.

Recently, Mee [96] has observed the efficiencies of the parameter estimates obtained from running a BB design, noticing the inefficiency with which 2-factor interactions are estimated. Recall that the efficiency of a parameter estimate $\widehat{\beta}$ is defined as $\sigma^2/(N\text{Var}(\widehat{\beta}))$. It is instructive to point out that in a one factor at a time design, no information can be gathered about the 2-factor interactions, i.e., they are not estimable. If we change several factors simultaneously from run to run, in a factorial fashion, there will be several cases where x_i and x_j will be varied together, and this adds information useful to estimate these effects. A BB design is in this sense somewhere in between an OFAT design and a factorial design; we should expect that the efficiency with which we estimate 2-factor interactions is not very good. Table 5.5 shows the efficiencies with which different BB designs estimate each of the parameters of relevance for process optimization. Note how for a PBIB, the efficiencies for the 2-factor interactions are not all the same, since the pairs of factors that are varied more often will have interaction effects estimated more precisely than the rest (actually, exactly twice as precise, since $m_1 = m_2/2$ in all cases). As it can be seen from the table, the efficiencies for estimating the 2-factor interactions are not good relative to the efficiencies of the other terms. It should be mentioned that BB proposed other designs for some values of k, but they require a considerable number of runs, so only the smallest BB design is listed. Mee's new BB-type designs [96] have in general much better efficiencies in

Table 5.5. Construction method and efficiencies of parameter estimates in selected Box-Behnken designs (source: from Mee [96])

				Efficiency		
k	Basic design	$N - n_0$	n_0	Linear	2 f. i.	Pure quad.
3	BIB: 2^2 3 times	12	3	0.533	0.267	0.247
4	BIB: 2^2 6 times	24	3	0.444	0.148	0.198
5	BIB: 2^2 10 times	40	6	0.348	0.087	0.19
6	PBIB: 2^3 6 times	48	6	0.444	0.148–0.296	0.19
7	BIB: 2^3 7 times	56	6	0.387	0.129	0.218
8	PBIB: 2^3 24 times	192	16	0.346	0.077–0.115	0.184
9	PBIB: 2^3 15 times	120	10	0.308	0.062–0.123	0.173
10	PBIB: 2^4 10 times	160	10	0.376	0.094-0.188	0.149

estimating the 2-factor interactions since they vary more factors at a time (for example, by using the columns of a 2^{5-1} design), but require in general a considerable number of runs.

5.5 Roquemore's (Hybrid) Designs

Roquemore [137] proposed an interesting set of designs for second order response surfaces. His goals when searching for these designs were:

1 First and third order moments must be zero to achieve orthogonality among most of the parameter estimates;

2 $[ii]$ and $[iijj]$ should be constant to equalize the variance of the linear and 2-factor interaction parameter estimates, respectively. This also promotes the rotatability of the design;

3 let $[iiii] = 3[iijj]$ to achieve rotatability;

4 design should be minimum in size.

The designs found meet most of these criteria (his designs are not always rotatable). They are based on ideas similar to a CCD. These designs have the following structure:

- At one level of variable k, a 2^{k-1} full factorial or a 2^{k-1-r} fraction of a 2^{k-1} in the remaining $k-1$ factors. This design may be rotated;
- At a second level of variable k, a cross polytope design (i.e., "axial points") for the remaining $k-1$ factors, possibly rotated;
- One or two axial points for variable k;
- A center point, if necessary, to avoid singularity of the moment matrix.

"Hybrid designs", as they are also called, can then be viewed as a central composite designs for the first $k-1$ factors, with one additional column for factor x_k and possible one or two more runs. The designs require 5 levels per factor, similarly than CCD's.

Interesting Roquemore designs are the 311A, the 416A, and the 628A. These are either rotatable or very near to rotatable, and very economical. The first number in the name refers to k, the next two numbers denote N (number of

runs) and the letter denotes the "version" number, as in some cases Roquemore proposed more than one design for the same k and N. The designs are as follows:

$$311\text{A} = \left[\begin{array}{rrr} x_1 & x_2 & x_3 \\ \hline 0 & 0 & 2 \\ 0 & 0 & -2 \\ -1.4142 & -1.4142 & 1 \\ 1.4142 & -1.4142 & 1 \\ -1.4142 & 1.4142 & 1 \\ 1.4142 & 1.4142 & 1 \\ 2 & 0 & -1 \\ -2 & 0 & -1 \\ 0 & 2 & -1 \\ 0 & -2 & -1 \\ 0 & 0 & 0 \end{array}\right]$$

$$416\text{A} = \left[\begin{array}{rrrr} x_1 & x_2 & x_3 & x_4 \\ \hline 0 & 0 & 0 & 1.7844 \\ 0 & 0 & 0 & -1.4945 \\ -1 & -1 & -1 & 0.6444 \\ 1 & -1 & -1 & 0.6444 \\ -1 & 1 & -1 & 0.6444 \\ 1 & 1 & -1 & 0.6444 \\ -1 & -1 & 1 & 0.6444 \\ 1 & -1 & 1 & 0.6444 \\ -1 & 1 & 1 & 0.6444 \\ 1 & 1 & 1 & 0.6444 \\ 1.6853 & 0 & 0 & -0.9075 \\ -1.6853 & 0 & 0 & -0.9075 \\ 0 & 1.6853 & 0 & -0.9075 \\ 0 & -1.6853 & 0 & -0.9075 \\ 0 & 0 & 1.6853 & -0.9075 \\ 0 & 0 & -1.6853 & -0.9075 \end{array}\right]$$

$$628\text{A} = \left[\begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ \hline 0 & 0 & 0 & 0 & 0 & 2.3094 \\ -1 & -1 & -1 & -1 & -1 & 0.5774 \\ 1 & 1 & -1 & -1 & -1 & 0.5774 \\ 1 & -1 & 1 & -1 & -1 & 0.5774 \\ -1 & 1 & 1 & -1 & -1 & 0.5774 \\ 1 & -1 & -1 & 1 & -1 & 0.5774 \\ -1 & 1 & -1 & 1 & -1 & 0.5774 \\ -1 & -1 & 1 & 1 & -1 & 0.5774 \\ 1 & 1 & 1 & 1 & -1 & 0.5774 \\ 1 & -1 & -1 & -1 & 1 & 0.5774 \\ -1 & 1 & -1 & -1 & 1 & 0.5774 \\ -1 & -1 & 1 & -1 & 1 & 0.5774 \\ 1 & 1 & 1 & -1 & 1 & 0.5774 \\ -1 & -1 & -1 & 1 & 1 & 0.5774 \\ 1 & 1 & -1 & 1 & 1 & 0.5774 \\ 1 & -1 & 1 & 1 & 1 & 0.5774 \\ -1 & 1 & 1 & 1 & 1 & 0.5774 \\ 2 & 0 & 0 & 0 & 0 & -1.1547 \\ -2 & 0 & 0 & 0 & 0 & -1.1547 \\ 0 & 2 & 0 & 0 & 0 & -1.1547 \\ 0 & -2 & 0 & 0 & 0 & -1.1547 \\ 0 & 0 & 2 & 0 & 0 & -1.1547 \\ 0 & 0 & -2 & 0 & 0 & -1.1547 \\ 0 & 0 & 0 & 2 & 0 & -1.1547 \\ 0 & 0 & 0 & -2 & 0 & -1.1547 \\ 0 & 0 & 0 & 0 & 2 & -1.1547 \\ 0 & 0 & 0 & 0 & -2 & -1.1547 \\ 0 & 0 & 0 & 0 & 0 & 0.0000 \end{array}\right]$$

The efficiencies with which each hybrid design estimates the effects of the second order model are shown on Table 5.6. As it can be seen, these widely vary from design to design in this class.

We describe next a method for evaluating the variance performance (hence, the efficiency) of a DOE in a more general form, by looking at the variance of the predictions over the experimental region.

Table 5.6. Efficiencies of parameter estimates for a second order model, selected Roquemore (Hybrid) designs

	Efficiency		
Design	**linear**	**2fi**	**pure quad.**
311A	1.4545	1.4545	1.5015–0.9697
416A	0.8550	0.5000	0.07–0.0465
628A	0.8571	0.5714	0.5714

5.6 Variance Dispersion Graphs

Variance Dispersion Graphs (VDG's, see [62]) are a graphical tool to evaluate the prediction variance properties of a design across the experimental region. To get a VDG for a design with associated matrix $\boldsymbol{X}$, we compute

$$V_r^M = \max_{\boldsymbol{x}'\boldsymbol{x}=r^2} \frac{N\text{Var}(\widehat{y}(\boldsymbol{x}))}{\sigma^2} = \max_{\boldsymbol{x}'\boldsymbol{x}=r^2} N\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$$

and

$$V_r^m = \min_{\boldsymbol{x}'\boldsymbol{x}=r^2} \frac{N\text{Var}(\widehat{y}(\boldsymbol{x}))}{\sigma^2} = \min_{\boldsymbol{x}'\boldsymbol{x}=r^2} N\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$$

the maximum and minimum scaled prediction variances, respectively, over the surface of spheres of radii r. Here, $\boldsymbol{x}_m$ denotes the vector $\boldsymbol{x}$ expanded in model form to match the columns of $\boldsymbol{X}$. VDG's also include a plot of the average prediction variance over spheres of radii r, defined as

$$\overline{V}_r = \frac{N}{\Psi\sigma^2}\int_{\boldsymbol{x}'\boldsymbol{x}=r^2} \text{Var}(\widehat{y}(\boldsymbol{x}))d\boldsymbol{x}.$$

where $\Psi = \int_{\boldsymbol{x}'\boldsymbol{x}=r^2} d\boldsymbol{x}$. A VDG plots V_r^M, V_r^m and $\overline{V}_r$ versus different radii r, typically going from $r = 0$ to $r = ||\boldsymbol{x}||_{max}$, the point in $\boldsymbol{X}$ farthest from the origin. This will generate three curves, which, evidently, can be graphed as a function of r regardless of the number of factors k, so VDG's are of greatest value when $k > 2$. Of the three lines, the one of utmost importance is V_r^M, the maximum prediction variance. V_r^m can be used to "break ties" when two designs have similar V_r^M.

The three lines on the VDG will always have values greater than 1.0, because

$$\frac{\text{Var}(\widehat{y}(\boldsymbol{x}))}{\sigma^2/N} \geq 1 \tag{5.9}$$

in other words, the VDG compares the prediction variance with the variance of an average of N observations obtained at the same point. This implies that designs with larger number of observations are compared more strictly, thus the scaled prediction variance is a way of asking larger designs to "justify themselves" [117]. It also allows to compare designs with different N. However, comparing designs with different number of runs in this way may not be a good idea. There is some debate about the value of looking at the scaled variance. This is a question that also arises in comparing designs using single number efficiencies, see Section 5.7 below. If two experimental designs with very different sample sizes are to be compared, we suggest to also look at the unscaled VDG. The unscaled graph will show the gain in precision of running a larger design and should be used to compare the two designs. The scaled VDG's can be used to determine how the prediction variances of each design compares to the variance of N replicates at a single point (see Equation (5.9).)

We can use VDGs to study the variance properties of some second order designs. Figure 5.1 shows the scaled VDG for three CCD designs[7] for $k = 6$ factors where the spherical choice of the axial distance $\alpha = \sqrt{k} = \sqrt{6}$ was used in all cases. The number of center points n_0 was varied from 1 to 5. As it can be seen, running only one center point results in high variance, and running more than 5 center points will not reduce the variance considerably further.

Figure 5.2 shows the scaled VDG[8] for another set of three CCDs, all for $k = 6$ factors and $n_0 = 5$ center runs (so $N = 49$). This figure illustrates the effect of the axial distance: the face-centered ($\alpha = 1$), spherical ($\alpha = \sqrt{k}$) and rotatable $\alpha = F^{1/4}$. Note how for a rotatable design, all three lines V_r^M, V_r^m, and $\overline{V}_r$, coincide.

In the literature, the face-centered CCD is evaluated with a different kind of VDG, the "cuboidal" VDG, in which the maximum, the minimum, and the average prediction variance is over squares of half side r. This is because it is argued it is a cuboidal design and "it is not natural to deal with surfaces over spheres when the design is cuboidal" [117, p. 409].

However, using a cuboidal VDG to evaluate a design is not a very sensible idea given the very different "shape" a hypercube in $k > 3$ dimensions can get which is against our geometric intuition. The distance to a corner in the $(-1,1)$

[7]The number of design points in each DOE is not dramatically different.
[8]All designs in this VDG have the same number of runs.

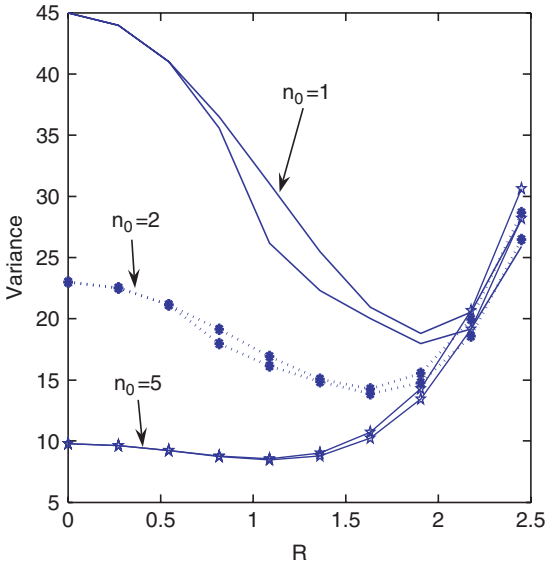


Figure 5.1. Effect of center points (n_0) on the prediction variance performance of a CCD for $k = 6$ factors. The spherical choice $\alpha = \sqrt{6}$ was used in all cases

hypercube is $\sqrt{k}$, which tends to infinity as $k \to \infty$. In contrast, the faces of the cube are always at a distance of 1.0, regardless of k. Thus, a cube in higher dimensions should be thought of as a "porcupine" [66, p. 41] with 2^k "spikes" that correspond to the corner (factorial) points. Computing the variance over points on such an irregular surface does not seem natural or advantageous for evaluating any DOE, since the points on which we will compute the min., max., and average $V(\widehat{y}(\boldsymbol{x}))$ can be at very different distances from the origin, with more drastic differences with increasing k. We will therefore use a "spherical" VDG regardless of the shape of the experimental region, since such a VDG will show the prediction performance of the design at points always separated *the same* distance r from the origin, as we move away from it, instead of showing the performance over sets of points with *different* or mixed distances from the origin.

The face centered CCD provides better prediction close to the origin (Figure 5.2). This design provides acceptable variance performance only close to the origin, with a maximum predicted variance that increases rapidly after $r > 0.5$. All the designs on the figure have $2^{6-1} = 64$ factorial runs at a distance of $\sqrt{6} \approx 2.45$, so the effect of the additional 12 axial runs is notable.

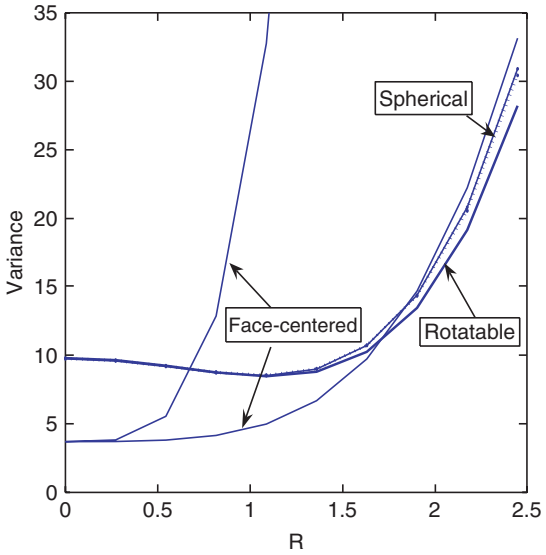


Figure 5.2. Effect of axial distance (α) on the prediction variance performance of a CCD for $k = 6$ factors and $n_0 = 6$ center points. The choices $\alpha = \{1, \sqrt{6}, F^{1/4}\}$ (face centered, spherical, and rotatable) are shown

The performance of the spherical CCD is better farther from the origin, a consequence of the effect of the axial distance. This is evidence against the use of face-centered CCD's, which are popular in practice mainly because they require only 3 levels.

Figure 5.3 shows a Roquemore 628A with 2 center points ($N = 30$), a CCD with $\alpha = \sqrt{6}$ and 3 center points ($N = 47$), and a BBD with $n_0 = 6$ ($N = 54$), the latter design table generated by the Minitab package. As it can be seen from the graph, the Roquemore design has an excellent performance relative to the others much larger designs. However, as it was mentioned before, one should not be carried away from looking at these scaled graphs. The unscaled variance of the Roquemore design will be considerable higher than that offered by the CCD. Perhaps the only clear thing from this graph is the poor variance performance of the Box-Behnken design.

Figure 5.4 shows a VDG for two designs for $k = 6$ with relative small number of runs: a Roquemore 628A with $n_0 = 2$ runs ($N = 30$) and a Small Composite Design (SCD, [49] which are similar to CCD's but utilize a resolution III factorial (or some other small fractional factorial) instead of a

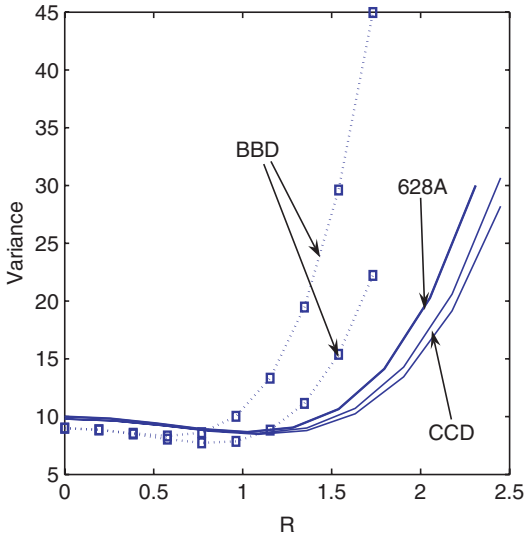


Figure 5.3. Some designs for $k = 6$ factors. A Box-Behnken with 6 center points ($N = 54$), a Roquemore 628 with 2 center runs ($N = 30$), and a spherical CCD with 3 center runs ($N = 47$)

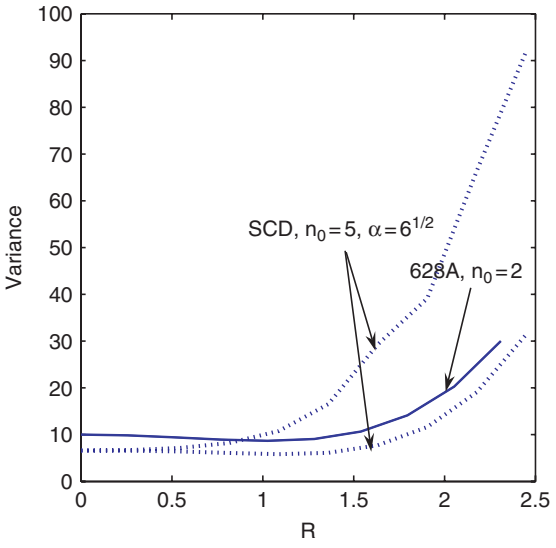


Figure 5.4. Economic designs for $k = 6$: a Small Composite Design ($N = 32$) and Roquemore's 628A ($N = 30$)

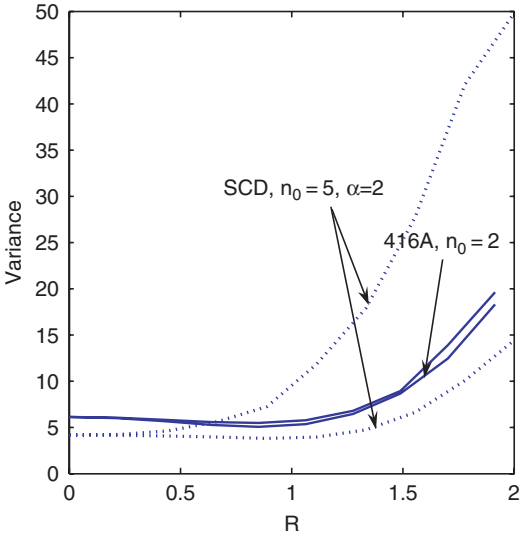


Figure 5.5. Economic designs for $k = 4$: a Small Composite Design ($N = 21$) and Roquemore's 416A ($N = 19$)

resolution V fraction. The SCD design shown has $n_0 = 5$ ($N = 32$), and the spherical axial value was chosen $\alpha = \sqrt{k}$. As it can be seen, only for the region next to the origin is the SCD preferable to the 628A, and not for much.

The last VDG on Figure 5.5 shows two designs for $k = 4$, a SCD with $n_0 = 5$ ($N = 21$), $\alpha = \sqrt{k}$, and Roquemore's 416A design with $n_0 = 3$ ($N = 19$). The conclusions are identical to those reached when looking at the previous figure for $k = 6$ factors.

5.7 D and G-Optimal Designs

Given the omnipresence of the $\boldsymbol{X}'\boldsymbol{X}$ matrix in the analysis of linear models, optimal design theory has seeked to modify this matrix so that it has good properties. In particular, optimal design theory has concentrated in making this matrix "small" in some sense. The problem is how to measure the size of a matrix. One of the most useful ways to do this is to use the determinant.

A D-optimal design is a design that maximizes $|\boldsymbol{X}'\boldsymbol{X}|$ for a given model (hence D stands for "determinant"). Recall that

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{\text{adj } (\boldsymbol{X}'\boldsymbol{X})}{|\boldsymbol{X}'\boldsymbol{X}|}$$

(see Appendix C) so maximizing the determinant will tend to make the entries in the inverse small. Because of this, D-optimal designs have:

1 small variances for the $\widehat{\beta}$'s;

2 small correlations between the $\widehat{\beta}$'s;

3 small variances for the predicted response.

D-optimal designs are very useful when we are quite certain of the model form, or when the experimental region is constrained in such a way that classical designs cannot be "fit" into the constrained region.

The main proponents of optimal design theory, Kiefer and Wolfowitz [80], developed a deep mathematical theory which we will only briefly mention here. An introductory account is the paper by St. John and Draper [148] and the more recent books by Spall [145] and Goos [64]. A classical but formal presentation is the book by Fedorov [54].

The basic idea in optimal design theory is to view a design as a probability measure ξ defined over the experimental region, that is:

$$\xi(\boldsymbol{x}) \geq 0 \;\; \boldsymbol{x} \in R, \quad \int_R \xi(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

An "approximate design[9]" ξ assigns to points in the experimental region a proportion or weight $\xi(\boldsymbol{x}) = w_i$. If $\xi(\boldsymbol{x}_i) = w_i = 0$ then there are no observations at point $\boldsymbol{x}_i$. To implement an approximate design in practice, $\boldsymbol{\xi}(\boldsymbol{x}_i)$ needs to be rounded in a corresponding exact design by n_i/N so that n_i observations out of the total are taken at x_i. In general, this discretization of the proportions cannot be done exactly as the values $\xi(\boldsymbol{x}_i)$ may be irrational, unless we let $N \to \infty$ (hence the name "asymptotic"). In Kiefer's approximate theory, a D-optimal design is defined as the design ξ^* such that

$$\xi^* = \arg\max_{\xi} \{|\boldsymbol{M}(\xi)|\}$$

where $\boldsymbol{M}(\xi)$ equals to

$$\sum_{i=1}^{n} w_i \boldsymbol{x}_{m,i} \boldsymbol{x}'_{m,i}$$

[9]Also called an asymptotic design. These names in the literature are used in contraposition to "exact" or "finite" designs that request a discrete number of observations be taken at each design point, i.e., a design $\xi(\boldsymbol{x})$ such that $N\xi(\boldsymbol{x})$ is an integer > 0.

and where $\boldsymbol{x}_{m,i}$ is a vector that contains the point $\boldsymbol{x}_i$ in "model form"[10]. In the exact case this reduces to $N\boldsymbol{M}(\boldsymbol{\xi}) = \boldsymbol{X}'\boldsymbol{X}$.

Box and Draper [20] criticize the determinant criterion[11] since the determinant, being a measure of the volume of the column vectors spanned by the matrix, will change with the "size" of the design, that is, the dispersion of the design points on the space of the controllable factors (note how "size" here is *not* the number of runs in the design but the dispersion of its points). Thus changes in the scale of the factors will have a dramatic effect on the determinant, particularly for large number of factors (high dimensions). This is certainly true, but as mentioned by Lucas [87], if all columns of a second order design are multiplied by a constant c, changing the scale of the k factors, then the determinant changes from $|\boldsymbol{X}'\boldsymbol{X}|$ to $c^{2k(k+2)}|\boldsymbol{X}'\boldsymbol{X}|$. Thus, although the determinant is not scale-invariant, a D-optimal design will remain so in a different scale, since the difference is a constant which plays no role in the optimization process. In other words, the D-*optimality* of a design is invariant to coding of the factors[12].

5.7.1 D-Optimal Design Algorithms

There is a large literature on D-optimal design construction. Most algorithms use an *exchange method*. These are based on a *candidate list* of points, frequently taken from a grid over the experimental region. The algorithm then starts to exchange points in the current design with those in the list, looking for improvements in $|\boldsymbol{X}'\boldsymbol{X}|$.

Cook and Nachtsheim [33] performed a benchmarking study of different exchange algorithms for finding exact D-optimal designs proposed up to 1980. They came to the not too surprising conclusion that the algorithms that took longer in searching for alternative designs would give in general more efficient designs. They concluded Mitchell's DETMAX [107] design to be the best overall, providing a balance between good efficiency and speed[13].

[10]That is, for a second order design in $k = 2$ factor, $\boldsymbol{x}_i = (x_1, x_2)'$ results in $\boldsymbol{x}_{m,i} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)'$.

[11]The critique is against all "alphabetic" criteria, as called perhaps in a derisive way by the authors.

[12]By the same token, efficiency comparisons shown later on, will not be affected either as long as the two designs being compared using ratios are scaled equally. Box and Draper proposed to use a coding convention that makes $\sum x_{iu}^2 = 1$ so that the designs one compares have the same "spread" over the controllable factors.

[13]There have been more recent approaches at D-optimality construction using genetic optimization algorithms that pose a great promise. For an introduction, see Heredia-Langner et al. [67].

A brief discussion of DETMAX and similar exchange algorithms will help to illustrate the complexities of the optimization problem.

Exchange methods were originally created to do the following: starting from a randomly generated N-point design, try:

a) adding an $N+1$ point such that $|\boldsymbol{X}'\boldsymbol{X}|$ increases the most. From an argument identical to the augmentation theorem (see 5.2 below), this point is that which maximizes $\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$ for $\boldsymbol{x} \in R$, i.e., for a point *anywhere* in the region of interest[14];

b) removing a point from the current design such that $|\boldsymbol{X}'\boldsymbol{X}|$ increases the most. By a similar argument as above, this is the point in the design at which $\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$ is *minimized*.

The two steps are tried until no further improvements are possible.

Step a) involves a hard, non-concave maximization problem, one that has many local maximums. Mitchell [107] followed an idea he attributed to Dykstra [52]: the maximization step in a) should be done over a grid of points. This is the origin of the "candidate list" of points L. Thus, all exchange methods do step a) for $\boldsymbol{x} \in L$ only.

It is clear then that, in general, the resulting design will not be *the* D-optimal design. It will approximate a D-optimal design the finer the grid is. Thus, the D-optimal design must be included in the candidate list for any such algorithm to find it. Finding good grids is in itself a problem, particularly on constrained regions R.

In some small, first order examples, however, it is possible to give the D-optimal design, as the next example shows.

Example. D-Optimal design for $k = 1$. If a linear model is going to be fit to a one factor data set, then $p = 2$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

[14]Here $\boldsymbol{x}$ are the coordinates of any point and $\boldsymbol{x}_m$ denotes the point expanded to match the columns of $\boldsymbol{X}$.

and

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

from which $|\boldsymbol{X}'\boldsymbol{X}| = n\sum x_i^2 - (\sum x_i)^2$. Thus. if we desire for example $N = 2$ runs and $-1 \leq x_i \leq 1$, then setting $x_1 = -1$ and $x_2 = 1$ gives the maximum value of $|\boldsymbol{X}'\boldsymbol{X}| = 4$. ■.

5.7.2 D-Efficiency

The *D-efficiency* of a design and model with given matrix $\boldsymbol{X}$ is defined as:

$$D_{eff} = \left(\frac{|\boldsymbol{X}'\boldsymbol{X}|}{\max|\boldsymbol{M}(\xi)|} \right)^{1/p}$$

where the denominator is the optimal value of the determinant for the approximate D-optima design. Thus, to know the D-efficiency of an exact design, we must know the efficiency of *the* approximate D-optimal design, otherwise, the denominator will not be the largest possible, and the ratio above will only be an overestimate of the actual D-efficiency, giving an *upper bound* on D_{eff}. Such upper bound can be computed by finding the best possible exact D-optimal design with an exchange algorithm and using its determinant in the denominator.

Because of these computational difficulties, for a given design with associated matrix $\boldsymbol{X}$, some D-optimal design programs (e.g., SAS OPTEX) report instead the quantity

$$D_{eff}^{(L)} = \frac{|\boldsymbol{X}'\boldsymbol{X}|^{1/p}}{N}.$$

This is a measure of D-optimality with respect to a hypothetical orthogonal design (which would be D-optimal[15]), but one that for the second order model is not achievable. It requires that the DOE be scaled (or re-scaled) such that $\sum_{u=1}^{N} x_{iu}^2 = 1$, which can be accomplished using Box and Draper's coding convention $x_{iu} = (\xi_{iu} - \overline{\xi}_i)/s_i$, where $s_i = \sqrt{\sum(\xi_{iu} - \overline{\xi}_i)^2}$.

The quantity $D_{eff}^{(L)}$ is a *lower bound* on the actual D-efficiency of a design, i.e., $D_{eff} \geq D_{eff}^{(L)}$. Although not a measure of absolute D-optimality, this is sometimes useful for comparing the relative efficiency of two or more designs with the same number of runs. However, the lower bound is sometimes not very "sharp".

[15]Note that for first order designs, orthogonal designs are D-optimal.

When comparing designs of *unequal* number of runs, the designs should be equally scaled (say, using the Box and Draper convention), because the determinant is a measure of volume, and this volume is determined by the scaling of the controllable factors. Simply reporting the determinant $|\boldsymbol{X}'\boldsymbol{X}|$ in cases of unequal number of runs is recommended as a measure of D-efficiency [64]. This is because the design that provides lowest $|\boldsymbol{X}'\boldsymbol{X}|$ will provide better parameter estimates.

In the example above for $k = 1$, the design is orthogonal, so $D_{eff}^{(L)} = D_{eff} = 1$.

5.7.3 D-Optimal Designs for First Order Models

Mitchell [108] investigated D-optimal designs for 1st order models using his DETMAX code. He points out that since Plackett Burman (PB) designs exist for N a multiple of 4, these are orthogonal and therefore variance optimal, as described in Chapter 3. If N is *not* a multiple of 4, there is no orthogonal design for a 1st order model, and a question of more theoretical than practical interest is what design to use if N cannot be rounded up to the next multiple of 4? By an application of the Design Augmentation Theorem 5.2, he shows that in a D-optimal design for a first order model on the hypercube $-1 \leq x_i \leq 1$ only consists of corner points. This facilitates considerably the construction of D-optimal designs, as the candidate points form a set with a finite number of elements. Mitchell conclusions are quite easy to follow:

- If $N = 4t$, use an orthogonal design (PB or 2^{k-r});
- if $N = 4t + 1$, add any corner point to an orthogonal design for N runs;
- if $N = 4t + 2$, add two runs to an orthogonal design for $N - 2$ runs such that the two points are as orthogonal as possible in the $\boldsymbol{X}$ matrix;
- if $N = 4t + 3$, remove any run to an orthogonal design for $N + 1$ runs.

In practice, it will be difficult to find a case where $4t + j$ runs are *mandatory* instead of simply running $4t$ or $4(t + 1)$ runs.

5.7.4 Design Augmentation

One of the most useful applications of D-optimal designs is when there is a need to "repair" or "augment" a design that has already been run. In this case,

we want to select the $(r+1)$th point in an existing design containing r points. The $r+1$ point is typically chosen from a candidate list of points such that it achieves D-optimality *given* the r-point design. The new design with $r+1$ points will not be D-optimal, in general. It turns out that finding the next point $\boldsymbol{x}_{r+1}$ is very easy[16], given a candidate set of points, as the following result shows.

THEOREM 5.2 *Given an r-point design and a model, both characterized by the matrix $\boldsymbol{X}_r$, the point $\boldsymbol{x}_{r+1}$ which maximizes $|\boldsymbol{X}'_{r+1}\boldsymbol{X}_{r+1}|$ is that for which $\boldsymbol{x}'_{r+1}(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1}\boldsymbol{x}_{r+1}$ is maximized throughout the candidate list of points.*

Proof. The proof is straightforward given the update formula for the determinant (which we do not prove here; see [1] for a proof):

$$|\boldsymbol{X}'_{r+1}\boldsymbol{X}_{r+1}| = |\boldsymbol{X}'_r\boldsymbol{X}'_r + \boldsymbol{x}_{r+1}\boldsymbol{x}_{r+1}| = |\boldsymbol{X}'_r\boldsymbol{X}_r|(1+\boldsymbol{x}'_{r+1}(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1}\boldsymbol{x}_{r+1})$$

so to maximize $|\boldsymbol{X}'_{r+1}\boldsymbol{X}_{r+1}|$ we maximize $\boldsymbol{x}'_{r+1}(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1}\boldsymbol{x}_{r+1}$.

Note how the quantity we need to maximize, $\boldsymbol{x}'_{r+1}(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1}\boldsymbol{x}_{r+1}$, is $\mathrm{Var}(\widehat{y}(\boldsymbol{x}_{r+1})/\sigma^2)$, *the scaled variance of the predicted response at the new point*. Thus, this result says that to "repair" a design we should look for the "weakest" spot and run an experiment there, a very intuitive recommendation. The procedure can be repeated several times sequentially, if more than just one point is desired.

Example. Consider a 2^2 factorial design used to fit a first order model in 2 factors. The next point is to be chosen from a 3^2 design. We then have $(\boldsymbol{X}'_r\boldsymbol{X}_r)^{-1} = \frac{1}{4}\boldsymbol{I}$, and if $\boldsymbol{x}' = (1, x_1, x_2)$, then $\mathrm{Var}(\widehat{y}(\boldsymbol{x}_{r+1}))/\sigma^2 = \frac{1}{4}(1 + x_1^2 + x_2^2)$. This equals to $3/4$ at any factorial (corner) point, $1/2$ at any midpoint of an edge, and to $1/4$ at the origin. Therefore, the sequential D-optimal design calls for replicating one of the corner points. ■.

5.7.5 G-Optimality and G-Efficiency

A good deal of the results on optimal design theory are due to the work by Kiefer and Wolfowitz [80]. One of the most useful results obtained by these authors concerns *G-optimality* and *G-efficiency*.

[16]In this section, $\boldsymbol{x}_{r+1}$ denotes the vector $\boldsymbol{x}$ expanded in model for to match the columns of $\boldsymbol{X}$.

A G-optimal design is a design described by matrix $\boldsymbol{X}$ such that

$$\min \max_{\boldsymbol{x} \in R} N\text{Var}(\widehat{y}(\boldsymbol{x})) = \min \max_{\boldsymbol{x} \in R} N\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m$$

where R is the experimental region and $\boldsymbol{x}_m$ indicates this vector $\boldsymbol{x}$ in model form. The outer minimization is over all possible designs. This minimax criterion assures the design obtained has no "weak spots" (from a prediction variance perspective) since its worst behavior over the experimental region is minimized.

It is easy to show (see Problem 13) that

$$\max_{\boldsymbol{x} \in R} N\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m \geq p \tag{5.10}$$

where p is the number of parameters in the model. Therefore, the G-efficiency of a design $\boldsymbol{X}$ is defined by

$$G_{eff} = \frac{p}{\max_{\boldsymbol{x} \in R} N\boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m}. \tag{5.11}$$

This has an obvious connection with the idea of leverage and the "Hat" matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Whenever $G_{eff} = 1$, all points in the design have the same "leverage" h_{ii}, where h_{ii} is the ith diagonal element of $\boldsymbol{H}$. Thus, the higher the G-efficiency of the design, the better the leverage of the points and the less likely influential observations will occur. However, one should recall that the h_{ii} terms are a measure of distance from each point i to the centroid of the points. G-optimal (and D-optimal) designs will tend to place the points on the boundaries of R, and therefore *all* points will tend to be about the same distance from the origin, resulting in a even leverage distribution[17].

5.7.6 Efficiencies of Classical Second Order Designs

Before the development of Variance Dispersion Graphs, it was customary to evaluate the efficiency of a design by looking at its G and D-efficiencies alone. We recommended to compute D and G-efficiencies in addition to the VDG when analyzing a design. Lucas [87] investigated the efficiencies of Central Composite Designs. Rather than repeating the tables in his paper, we can summarize his conclusions:

[17]However, Lucas [90] has argued that G-efficiency is usually the best single performance measure of an experimental design.

- $|\boldsymbol{X}'\boldsymbol{X}|$ is an increasing function of α in a CCD. Thus, for a spherical region, the optimum value of the axial distance is $\alpha = \sqrt{k}$ and for a hypercube with side equal to two, the best choice is $\alpha = 1$;
- CCD's can be constructed that are close to being D-optimal for $k \leq 5$. These CCD's have a resolution V or greater 2^{k-r} factorial for $n \geq 6$;
- As k increases, the type of CCD's mentioned above have a D-efficiency that gets farther from being D-optimal, although the decrease in efficiency is not very dramatic. For $k = 11$ it is only down to 0.83;
- Center points almost always decrease the efficiency of a CCD, although there are other good reasons to include them, as described before.
- Designs with resolution III or lower are never D-optimal.

In a subsequent paper, Lucas [88] studied the D and G efficiencies of CCDs, Box-Behnken designs, and other designs used for second order polynomial models, both under a spherical and under a hypercube (in contrast with VDGs, efficiency computations should be modified depending on the shape of the experimental region, as the optimization problem has different constraint types). His conclusions were that BBD's and CCD's are very efficient, and that "while more efficient designs are possible, they either remain to be discovered or they require significantly more experimental points. Classical designs will continue to be used in most applications" [88]. That CCD's and BBD's have been continuously used is certainly true, but recent work indicates that BBDs are not that good designs as once thought from the variance/efficiency point of view (see Sections 5.4 and 5.6).

The designs by Roquemore (see Section 5.5) indicate that there *is* room for improvement in second order experimental designs; not everything has been written about this subject. Some of these designs are extremely efficient, economical, and in some cases are rotatable.

5.7.7 Kiefer-Wolfowitz Theorem

A fundamental result was proved by Kiefer and Wolfowitz. Their equivalence theorem indicates that *for approximate designs*, D-optimality and G-optimality are equivalent conditions. Unfortunately, this does not happen for *exact* designs, otherwise checking D-optimality would be as easy as checking

G-optimality. In many cases, however, we have that $G_{eff} < D_{eff}$, although there are exceptions to this.

5.7.8 Good Practice when using D-Optimal Designs: Degrees of Freedom and Variance Inflation Factors

D-optimal designs are of great value when the experimental region is constrained in a non-standard way, when the model form is quite well known, and when repairing a bad design or augmenting a design with a few more runs is desired. Good practice when using a D-optimal or sequential D-optimal design is to also consider the degrees of freedom available to do model tests and to look at the $\boldsymbol{X}'\boldsymbol{X}$ matrix for problems.

In practice, it is suggested that the resulting designs should have:

- $n - p \geq 10$ (or so) to have enough degrees of freedom to obtain a good estimate of variance and reduce leverage;
- $m > p$ to allow testing lack of fit[18] (recall that m is the number of distinct design points);
- $n > m$ to have enough degrees of freedom for "pure error".

In practice, these three inequalities can be entered in a D-optimal algorithm as additional constraints under which the determinant objective is maximized.

Another issue to consider relates to lack of orthogonality, what in regression is called multicollinearity. This is reflected in the degree of linear association among the columns of the $\boldsymbol{X}$ matrix. For a second order model, complete orthogonality is impossible, but trying to get a design as orthogonal as possible is a good idea since the parameter estimates will then have low variance and covariances. A measure of multicollinearity are the **Variance Inflation Factors.** These are simply the diagonal elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix when written in correlation form. The jth variance inflation factor, VIF_j, is simply the scaled variance of the estimate of the jth coefficient in the model. There are various rules of thumb about what should be considered a high VIF, and therefore, evidence of multicollinearity. For a second order model, it is probably wise to say that if $\text{VIF}_j > 10$, then there is multicollinearity in the design,

[18]This recommendation is related to the "bias vs. variance" debate discussed in Chapter 8.

which should be fixed. When this problem is detected, we need to "fix" the design with additional runs that decrease the linear dependencies. The D criterion may be useful for such purpose, since $|\boldsymbol{X}'\boldsymbol{X}|$ can be maximized with respect to the additional runs subject to constraints that limit the VIFs.

If there is no chance of augmenting the design and this has to be used "as is", then the analysis should take into consideration the multicollinearity. Ridge regression techniques is one way of doing this, see [111].

5.8 Mixture Experiments

A different kind of experimental design where second order and higher order models are often fit is a mixture design. Since we will refer a few times in subsequent chapters to this kind of experiments, we provide a brief overview in this section for completeness. For a thorough discussion of mixture designs and analysis, see the book by Cornell [34].

The distinctive feature of a mixture is that the controllable factors are the **ingredients** of a mixture. The ingredients must add up to 100%, and the response or responses of interest are modelled as a function of these proportions[19]. The factors are then called the **components** of the mixture. If x_i denotes the ith component and q the number of components in the mixture, the mixture constraints are:

$$\sum_{i=1}^{q} x_i = 1, \qquad 0 \leq x_i \leq 1,\ i = 1, \ldots, q. \tag{5.12}$$

This constraint modifies the space of controllable factors from $\mathbb{R}^q$ to a $q-1$ dimensional **simplex** (see Figure 5.6). For $q = 2$ components, the simplex is simply a line; in one extreme $x_1 = 1$ and $x_2 = 0$; at the other extreme, $x_1 = 0$ and $x_2 = 1$. All points such that $x_1 + x_2 = 1$ are in the simplex. In general, all points $\boldsymbol{x}$ such that (5.12) is true are inside the simplex.

Examples of mixture experiments abound. Some examples are alloys, cement, plastics, gasoline, many examples in the food industry, soap, detergent, and even some semiconductor manufacturing processes, such as etching. Sometimes a hidden ingredient, typically water, is what makes all components add up to 100%. Water is sometimes thought to be an "inert" component, and the experiment is run (incorrectly) as a factorial without considering the "water" component when in fact this is a mixture experiment. The main

[19] Strictly speaking, the ingredients need only add up to a fixed proportion, not necessarily 1.

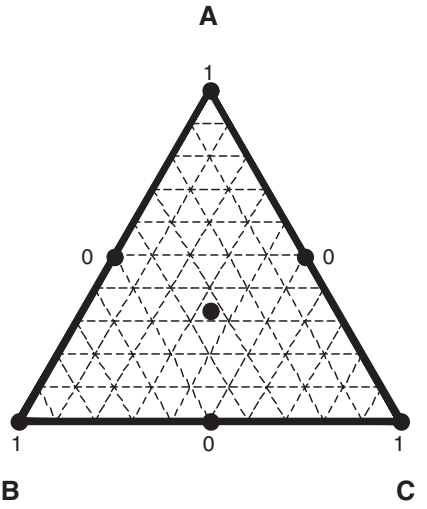


Figure 5.6. The simplex experimental region for $q = 3$ components. The vertices have coordinates $x_i = 1, x_j = 0$ for $j \neq i$

difference between a factorial and a mixture experiment is that in a factorial, all factors can take values independent of each other, whereas in a mixture the values of the components are constrained by (5.12).

5.8.1 Simplex Designs and Associated Polynomial Models

The simplest instance of a mixture experimental design is a **Simplex Lattice Design** in which there is a uniformly spaced set of points on the simplex. A $\{q, m\}$ simplex lattice design is simply a design in which each of the q components takes proportion values at the $m + 1$ equally-spaced values:

$$x_i = 0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m} = 1, \quad \text{for } i = 1, 2, \ldots, q,$$

and all possible *mixture* combinations of these proportions are used in the design, i.e., the combinations must satisfy (5.12).

Example. a {3,2} lattice design. We have that the values on the lattice are $x_i = 0, 1/2$ and 1 for $i = 1, 2, 3$, and therefore the design consists of the points

$$(x_1, x_2, x_3) = \{\underbrace{(1,0,0), (0,1,0), (0,0,1)}_{\text{“pure blends”}}, \underbrace{\left(\frac{1}{2}, \frac{1}{2}, 0\right), \left(\frac{1}{2}, 0, \frac{1}{2}\right), \left(0, \frac{1}{2}, \frac{1}{2}\right)}_{\text{“binary blends”}}\}$$

Table 5.7. A $\{3, 2\}$ simplex lattice design

Run	A	B	C
1	1	0	0
2	0.5	0.5	0
3	0.5	0	0.5
4	0	1	0
5	0	0.5	0.5
6	0	0	1

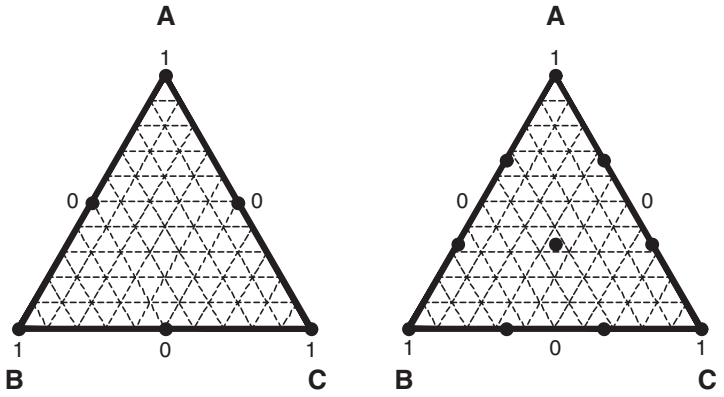


Figure 5.7. Left: a $\{3, 2\}$ and (right) a $\{3, 3\}$ simplex lattice design

The corner points are often called the "pure blends", although they are actually not blends, and the 3 last points listed above do mix two of the three components and hence their name. Table 5.7 shows the resulting design. Figure 5.7 shows the corresponding design, together with a $\{3, 3\}$ simplex lattice design. Notice how the $\{3, 2\}$ design does not place any design point in the center, whereas the $\{3, 3\}$ design places one run in the simplex centroid $(1/3, 1/3, 1/3)$. ■

The total number of points in a $\{q, m\}$ simplex lattice design is

$$\left(\begin{array}{c} q+m-1 \\ m \end{array} \right).$$

A form of regression model that can be fitted using a $\{q, m\}$ simplex lattice design is the **canonical** or **Scheffé form** of a mixture polynomial model.

An important feature of every mixture design from the model estimation point of view is that the mixture constraint introduces a perfect linear relation in the $\boldsymbol{X}$ of the design. The canonical form is derived by avoiding such linear dependencies in such a way that the final model is estimable.

For example, for $q = 2$ components, the first order Scheffé polynomial model is

$$\begin{aligned} E[y] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &= \beta_0(x_1 + x_2) + \beta_1 x_1 + \beta_2 x_2 \\ &= (\beta_0 + \beta_1)x_1 + (\beta_0 + \beta_2)x_2 \\ &= \beta_1^* x_1 + \beta_2^* x_2 \end{aligned}$$

where the first step follows due to the mixture constraint (5.12). The procedure has hence removed the intercept β_0, and thanks to this, the parameters β_1^* and β_2^* are estimable.

In general, with a $\{q, m\}$ simplex lattice design one can fit a Scheffé polynomial of order m. For q components, the canonical forms of the mixture models (with the * removed from the parameter names) are as follows.
Linear model:

$$E[y] = \sum_{i=1}^{q} \beta_i x_i$$

(with q terms). This model is adequate if blending is additive.
Quadratic model:

$$E[y] = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{j>i}^{q} \beta_{ij} x_i x_j$$

which has $q + \frac{q(q-1)}{2} = \frac{q(q+1)}{2}$ terms.
The "Full cubic" model is

$$\begin{aligned} E[y] &= \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{j>i}^{q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \sum_{j>i}^{q} \delta_{ij} x_i x_j (x_i - x_j) \\ &\quad + \sum_{i=1}^{q} \sum_{j>i}^{q} \sum_{k>j}^{q} \beta_{ijk} x_i x_j x_k \end{aligned}$$

which has $\frac{q(q+1)(q+2)}{26}$ terms. In this model, the pure cubic terms are also removed to avoid singularity of the $\boldsymbol{X}'\boldsymbol{X}$ matrix.

Finally, the "Special cubic" model is:

$$E[y] = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \sum_{j>i}^{q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \sum_{j>i}^{q} \sum_{k>j}^{q} \beta_{ijk} x_i x_j x_k$$

which does not contain any $x_i^2 x_j$ term and has $\frac{q(q^2+5)}{6}$ terms.

In general, the number of terms in an m-order Scheffé polynomial model is

$$\left(\begin{array}{c} q+m-1 \\ m \end{array} \right).$$

which is exactly equal to the number of points in an $\{q, m\}$ simplex lattice design.

In all of these models above, the coefficients β_i do not represent "slopes" or derivatives of y as x_i changes (as in factorial–non-mixture– experiments), but instead they represent the *height* of the response surface at the "pure blend" points ($x_i = 1, x_j = 0, j \neq i$). The $\beta_{ij} x_i x_j$ terms represent the excess response of the quadratic model over the linear model. These terms allow to model curvature along the *edges* of the simplex, but not in its interior. If $\beta_{ij} > 0$ the blending is said to be *synergistic* because as the two components i and j are introduced in the mix the response increases. If $\beta_{ij} < 0$ the blending of the components i and j is said to be *antagonistic* because as these two components are introduced in the mix the response goes down.

Simplex lattice designs can be augmented with center points (at the centroid) and/or with "axial" points, so that more interior points are run. Other standard designs for mixtures have been proposed, such as Simplex centroid designs (useful to fit the "special cubic" model) and Becker designs. Constraints in mixture designs are very common. If the constraints are such that the feasible region is a simplex, then the standard designs can be applied in the reduced region. When this is not the case, D-optimal designs are very useful. An issue in mixture designs is multicollinearity. Even after eliminating the simplex constraints from the model, there is always significant multicollinearity. Some analysts indicate that VIF values between 50 and 100 are not to be considered uncommon in mixture experiments.

The analysis of variance for mixture experiments proceeds along similar lines than in non-mixture experiments. However, there are some important differences. The null hypothesis: $H_o : \beta_1 = \beta_2 = \cdots = \beta_q = \beta$ tests for

no linear blend, i.e., the hypothesis says the response is just a plane parallel to the simplex region. Effects are also defined differently than in factorial experiments, given the mixture constraint. The effect of a component i is defined as

$$E_i = \widehat{\beta}_i - \frac{\sum_{j \neq i} \widehat{\beta}_j}{q-1}$$

and measures the effect of increasing x_i while keeping the relative proportions of all other components constant.

Experiments in which a subset of factors are components of a mixture and some other factors are not, for example some other factors are process variables, are also common in practice. The design and analysis of these experiments is treated by Cornell [34].

5.9 Problems

1 Show that using $\alpha = \sqrt{k}$ results in a rotatable CCD for $k = 2$ and $k = 4$, but not for $k = 3$.

2 What is the effect of varying the λ_4 parameter when using a CCD? How do you vary this parameter in practice?

3 Consider the following experimental design in two factors that is to be used to fit a second order polynomial model:

x_1	x_2
$-g$	$-g$
$-g$	0
$-g$	g
0	$-g$
0	0
0	g
g	$-g$
g	0
g	g

where g is some constant. Is this experiment rotatable? Why?

4 Suppose we want to use a CCD for $k = 6$ factors, where the factorial portion of the experiment is a 2^{6-1} fractional factorial.

a) Find the axial distance (α) for a rotatable CCD design.

b) Find the number of center points to achieve (near) uniform precision.

c) Find the number of center points that will give an "orthogonal" CCD.

d) Suppose the CCD is going to be run in two blocks (axial and factorial, with a single block in the factorial part). Allocate center runs to each block to allow for the design to block orthogonally while at the same time obtaining (near) uniform-precision, and near rotatability.

5 Consider a Box-Behnken design for 3 factors. Prove that this is *not* a rotatable design.

6 In this exercise you will compare, for $2 \leq k \leq 10$ "spherical" CCD's (i.e., $\alpha = \sqrt{k}$) and rotatable CCD's. Specifically:

a) which type of DOE places the points farthest from the origin?

b) for which values of k are spherical designs rotatable?. For the rotatable CCD's, compute α using the smallest 2^{k-r} fraction of resolution at least V.

7 Consider a Box-Behnken design for 5 factors. Prove that this is *not* a rotatable design.

8 Show that Hybrid designs 311A, 416A+one center point, and 628A are all rotatable.

9 Show that for a first order model, D-optimality implies rotatability.

10 Suppose we have a 2^3 factorial design used to fit a first order model in 3 variables. We want to add one more point using the idea of sequential D-optimality. The list of candidate points consists of all the points in a rotatable CCD.

a) Give the next D-optimal point given the current design.

b) Give the determinant of the resulting $\boldsymbol{X}'_{r+1}\boldsymbol{X}_{r+1}$ matrix using the update formula for the determinant.

11 Suppose we wish to fit a full quadratic model in $k = 4$ factors. A 20-run D-optimal design for this problem is shown below in the natural variables.

a) Is the leverage distribution appropriate? Why?

ξ_1	ξ_2	ξ_3	ξ_4
0.64	340	220	480
0.96	510	220	480
0.96	340	330	480
0.64	510	330	720
0.64	510	220	600
0.80	425	220	720
0.64	510	330	480
0.96	340	330	720
0.64	340	330	480
0.80	510	275	720
0.64	340	220	720
0.80	425	220	480
0.80	340	275	600
0.64	425	275	600
0.96	340	275	720
0.96	340	220	600
0.96	425	275	480
0.96	510	220	720
0.64	340	330	720
0.96	510	330	600

b) Find the VIF's of this design. (Hint: to find the VIF's, use the "unit length" coding convention:

$$x_{iu} = \frac{\xi_{iu} - \overline{\xi}_i}{\sqrt{\sum_{u=1}^{n}(\xi_{iu} - \overline{\xi}_i)^2}}$$

which gives $\sum_{u=1}^{n} x_{iu} = 0$ and $\sum_{u=1}^{n} x_{iu}^2 = 1.0$. If you do this, the $\boldsymbol{X'X}$ matrix will be in *correlation or moment form*. Then the diagonal elements of $(\boldsymbol{X'X})^{-1}$ will give you the VIF's.)

12 Compare the D-Optimal design in the previous problem with Roquemore's 416A design with 4 center runs added ($n = 20$ runs) based on leverage (h_{ii}) and orthogonality (VIF's). (Hint: use the same "unit length" coding convention as above. Otherwise, your comparison will not be fair.)

13 Show that expressions (5.10) and (5.11) for the G-efficiency of a design are true.

14 Consider a 2^2 design with center points used to fit a first order model (i.e., intercept and main effects only). Show that if the number of center points, n_0, is greater than zero, then the resulting design is not G-optimal.

15 In a *2 component* mixture experiment, car mileage figures were recorded based on using only each of the individual fuels A and B as well as using a 50%:50% blend of the two fuels A and B. Consider the following three separate experimental studies with mileages as follows:

Study no.	**A only**	**B only**	**A and B**
1	17	10	15
2	6	6	4
3	9	12	12

In which of the experimental studies are the fuels *synergistic*? In which are the fuels *antagonistic*?

16 One way of using a regular experimental design in a mixture situation is to assign the factors to ratios of components. Suppose that in a 3-component experiment you decide to use the component ratios $r_1 = x_2/x_1$ and $r_2 = x_3/x_1$ as your two independent factors. A rotatable CCD for 2 factors with 3 center runs is going to be run on these ratios.

a) For each CCD design point, identify the corresponding point on the simplex for the mixture components (x_1, x_2, x_3). Assume the lowest ratio possible is zero.

b) Suppose the same CCD design as above is to be used. In addition, suppose we only wish to investigate mixtures where x_1 (first component) is lower than 0.5. What is a good coding scheme that will achieve this? (Hint: what is the largest allowable ratio $(\max(r_1, r_2)$ you should allow?)

17 Derive the Scheffé second order (quadratic) polynomial model by applying (5.12) and $x_i^2 = x_i(1 - \sum_{j \neq i} x_j)$ to a standard (non-mixture) second order polynomial and simplifying.

18 If there are 8 components in a mixture experiment, what is the number of terms in a: a) quadratic model, b) special cubic model?

19 What DOE's will allow us to fit the two models in the previous question?

PART III

STATISTICAL INFERENCE IN PROCESS OPTIMIZATION

Chapter 6

STATISTICAL INFERENCE IN FIRST ORDER RSM OPTIMIZATION

The history of science counsels us to hold our mathematical generalizations regarding the universe lightly. But it also encourages us to anticipate that a first approximation may be valuable in suggesting a closer fidelity to the observables of nature.

—Eric Temple Bell (1883–1960)

All of the standard inferences in RSM as presented in previous chapters are based on point estimators which have sampling, or experimental, variability. Assuming a classical or frequentist point of view, every quantity computed based on experimental data is subject to sampling variability and is therefore a random quantity itself. As Draper [48] pointed out, one should not expect precise conclusions when using mathematical optimization techniques based on data subject to large errors. This comment applies to every technique previously discussed, namely, the steepest ascent/descent direction, eigenvalues of the quadratic matrix and point estimators of the stationary or optimal points in quadratic (second order) optimization for both canonical and ridge analysis. It also applies to more sophisticated mathematical programming techniques. In the RSM literature, there has been an over-emphasis on using different types of such mathematical techniques which neglect the main statistical issue that arises from random data: if the experiment is repeated and new models fitted, the parameters (or even the response model form) may change, and this will necessarily result in a different optimal solution.

More specifically, for steepest ascent, observed response drops in the steepest ascent direction may be due to noise and not really because the mean response is decreasing. This indicates the need for a statistical stopping rule for steepest ascent/descent that distinguishes between "noise" and the real "signal" of a response that is decreasing. In addition, the steepest ascent/descent direction may not be the best direction for searching. Some recent work has discussed, based on statistical considerations, how to search for an ascent direction other than steepest ascent. These are topics we explore in this chapter, namely, cases where statistical issues arise in first order RSM optimization. Chapter 7 will discuss in turn how to consider sampling variability in optimization techniques for second order models.

6.1 Stopping Rules for Steepest Ascent

As we commented at the end of Chapter 2, an aspect of the Box-Wilson methodology which was not well defined was when to stop a steepest ascent/descent search. If we are maximizing a response (say), a drop in the response may be due to "noise", not due to a real decrease of the mean response. We wish to find a rule that can determine when to stop in such a way that we neither miss the true maximum over the steepest ascent direction due to an early stoppage nor do we stop experimenting several steps after the maximum occurs, which will result in waisted experiments. In this section we describe two methods that have been proposed to answer this question.

We assume in this section the step size along the search direction has already been selected according to what was described in Chapter 2. It is important to point out that the stopping rules described in this section apply to any given search direction, not only the steepest ascent/descent one. Hence, we will simply refer to a generic and given "search direction" in this section.

Let t be the number of steps along the search direction on which we wish to conduct an experiment. The response at step t on such direction can be modeled as

$$y(t) = \eta(t) + \varepsilon, \quad \varepsilon \sim (0, \sigma^2)$$

where we assume for the moment that the errors have a zero mean, constant variance distribution. An underlying assumption is that in the search direction the unknown function has an unimodal mean $\eta(t)$. Suppose we wish to maximize the mean response, without loss of generality.

Stopping rules for searching along a direction can be quite simple, for example we can apply a:

- "Classical" rule: stop at the first step t_{stop} such that $y(t_{stop}) < y(t_{stop} - 1)$.
- 2-in-a-row rule: stop at the first step t_{stop} such that $y(t_{stop}) < y(t_{stop}-1) < y(t_{stop} - 2)$.
- 3-in-a-row rule: stop at the first step t_{stop} such that $y(t_{stop}) < y(t_{stop}-1) < y(t_{stop} - 2) < y(t_{stop} - 3)$.
- etc.

A stopping rule will return (for a maximization problem, like in steepest ascent) the value t_{opt}, where

$$y(t_{opt}) = \max_{t=1,2,\ldots,t_{stop}} \{y(t)\}. \tag{6.1}$$

We now describe two formal stopping rules.

6.1.1 Myers and Khuri Stopping Rule

Myers and Khuri (MK, [115]) proposed the following sequential test of hypothesis approach for stopping experimentation. Without loss of generality, assume a maximization (ascent[1]) problem.

Let $i = 0$ (the index i represents the number of steps since we last start testing). Let $j = 1$. *If there is a drop in the observed response*, say from step n_j to $n_j + 1$ (i.e., $y(n_j) > y(n_j + 1)$, thus n_j is the step number just before the jth observed drop[2]) we set $i \leftarrow i + 1$, and test:

$$H_0^{(j)} : \eta(t) \geq m_o^{(i)} \quad \text{i.e., the process mean is increasing at step } i$$
$$H_1^{(j)} : \eta(t) \leq m_1^{(i)} \quad \text{i.e., the process mean is decreasing at step } i$$

where $t > n_j$, $m_i^{(i)} - m_o^{(i)} = \Delta < 0$ and $m_o^{(i)} = \eta(n_j)$. The MK test is:

1 Accept $H_0^{(j)}$ if $y(n_j + i) - y(n_j) \geq b > 0$. Make $i = 0$. In this case, we keep experimenting along the (ascent) direction, and we repeat the test only when a new drop is observed in which case we make $j := j + 1$;

[1]Myers and Khuri derived their procedure for the steepest ascent direction, but evidently, their procedure applies without modification to any ascent direction.

[2]Note how we keep track of both the drop number j and when –in which step– each drop occurred, n_j.

2 Reject $H_0^{(j)}$ if $y(n_j+i)-y(n_j) \leq a < 0$. In this case we stop experimenting along the (ascent) direction and return t_{opt} using (6.1);

3 Keep experimenting along the (ascent) direction if $a < y(n_j+i)-y(n_j) < b$. This means make $i := i + 1$, and keep testing if the response at $n_j + i$ is significantly lower than that at n_j.

The bounds a and b derived by MK are elegantly simple:

$$a = -b = \Phi^{-1}\left(\frac{1}{2\kappa}\right)\sigma\sqrt{2}$$

where $a \leq 0$ for $\kappa \geq 1$ and Φ^{-1} denotes the inverse of the standard normal CDF. These bounds were derived by solving

$$\min\ E\left[t_{stop} | \bigcap_{i=1}^{\infty} H_1^{(i)}\right]$$

(minimize the number of experiments until stopping given the response was decreasing) subject to

$$E\left[t_{stop} | \bigcap_{i=1}^{\infty} H_0^{(i)}\right] \geq \kappa$$

which safeguards against missing the optimum, that is, stopping when the mean response was actually increasing. The value of κ is an *a priori* estimate (a lower bound) on the number of steps until the maximum is observed along the search (ascent) direction. Thus this is an estimate of what we wish to find, t_{opt}, and therefore this is a weakness of this method. We will mention the effect of selecting an erroneous value of κ in the performance of the procedure below.

Notice how the comparisons are made between the observation right before a drop was observed, $y(n_j)$, and the subsequent observed response values $y(n_j + i)$, hence the need to keep two subscripts, one for the *drop* number and one for the *step* number since the last drop. Two simple illustrations will suffice to explain the algorithmic description.

Example. Myers-Khuri stopping rule for steepest ascent illustration. Consider Figure 6.1. The figure shows two typical behaviors of the MK rule for some given values of a and b. Let us suppose the search direction is the steepest ascent direction. Consider first the graph on the left in Figure 6.1. The figure shows the label "NT" whenever no test was performed. Since the first drop

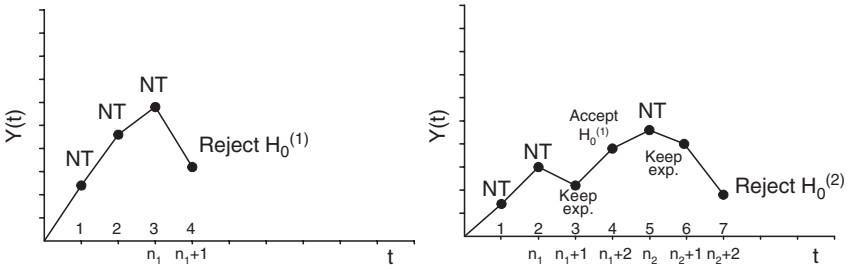


Figure 6.1. Two examples illustrating the Myers-Khuri stopping rule. On the left, $t_{stop} = 4$ and $t_{opt} = 3$; on the right, $t_{stop} = 7$ and $t_{opt} = 5$

occurs from step 3 to step 4, $n_1 = 3$ and $n_1 + 1 = 4$. Let us suppose a is such that the difference $y(n_1 + 1) - y(n_1) < a$ (recall a is negative), so we would reject $H_0^{(1)}$ at step $t_{stop} = 4$ and return $t_{opt} = 3$, the step at which the highest response was observed. Now consider the graph on the right of the Figure. In this case, no tests were conducted at steps 1 and 2, but a drop occurred from step $n_1 = 2$ to step $n_1 + 1 = 3$. The difference $y(n_1 + 1) - y(n_1)$, however, was in the interval (a, b), so we cannot conclude if the drop was significant or not. Thus, we keep experimenting along the steepest ascent direction. At step $4 = n_1 + 2$, the difference $y(n_1 + 2) - y(n_1) > b$ (note we compare against $y(n_1)$, since we could not conclude earlier if the first drop was significant or not), so we accept the null hypothesis that the mean response was really increasing (accept $H_0^{(1)}$). We set $i = 0$ and keep experimenting, waiting to observe another drop in the response. No test is conducted at step 5 since the response did not drop. However, the second ($j = 2$) drop was observed from step $5 = n_2$ to step $6 = n_2 + 1$. The difference $y(n_2 + 1) - y(n_2)$ is between the (a, b) range, so we keep experimenting along the steepest ascent direction and keep testing, making $i = 2$. The difference $y(n_2 + 2) - y(n_2)$ (note we compare against $y(n_2)$) is now less than a, so we reject $H_0^{(2)}$ and stop at $t_{stop} = 7$, returning $t_{opt} = 5$. ■

Simulation experience with the Myers-Khuri stopping procedure ([37], [102]) indicates that it has a good performance compared with the n-in-row rules mentioned before. In simulated parabolic functions, the MK rule stops after the optimum, but not too far after it occurs on the steepest ascent/descent direction. The "first drop" and "two drops in a row" stopping rules tended to stop earlier than the optimum in a large number of cases, *so they should not be*

used. Waiting 3 drops in a row before stopping wastes experiments (stops too many steps after the optimum), but because of this, the optimum is not missed.

The Myers-Khuri rule is sensitive to the normality assumption (made when computing the bounds a and b). The effect of using $\kappa > t_{opt}$ is more moderate than if $\kappa < t_{opt}$, thus this rule should be used with a non-conservative *a priori* estimate of where we think the optimum might lie along the search direction.

6.1.2 Recursive Parabolic Stopping Rules

In addition to the rule by Myers and Khuri, a rule by del Castillo [37] and a rule by Miro and del Castillo [102] have been proposed for stopping a search based on experiments[3]. The Recursive Parabolic Rule (RPR) [37] fits a parabola (i.e., a second order model) to the sequence of observed responses in the search direction. It recursively updates the estimate of the second order coefficient and tests if the first derivative is negative. When this is true, that is evidence a maximum has been reached and the search is stopped.

The performance of the RPR was only studied in [37] when the response function was quadratic with additive noise. Although classical RSM uses second order models to obtain a local approximation of the true function, during steepest ascent searches significant changes are made to the controllable factors and the quadratic approximation may not be valid over a large region of the factor space. Furthermore, the intercept and first order term carried over from the previous factorial design may also need to be changed in order to give a better local approximation of the true function. For these reasons, Miro et al. [102] proposed several enhancements to the RPR rule to make it robust with respect to non-quadratic behavior.

In this section we first present the RPR rule and then the enhancements proposed to it by Miro et al. [102] are discussed.

Recursive Parabolic Rule.

The Recursive Parabolic Rule[4] (**RPR**, [37]) assumes that the mean response along the search direction can be modeled by a quadratic polynomial of the

[3]As in the previous section, these alternative stopping rules are applicable on any given direction, not only the steepest ascent/descent direction. We will refer to their application along the steepest ascent/descent direction in this section.

[4]This section is based on reference [102].

form:

$$\eta(t) = \theta_0 + \theta_1 t + \theta_2 t^2 \ , \tag{6.2}$$

where t is the step counter of the search. The values of the true response $\eta(t)$ cannot be directly observed:

$$Y(t) = \eta(t) + \epsilon_t = \theta_0 + \theta_1 t + \theta_2 t^2 + \epsilon_t \ , \tag{6.3}$$

where $Y(t)$ are the observed values during the search and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ is a sequence of i.i.d. random variables.

Under the assumptions in model (6.2) the point of maximum mean response is given by the value of t such that:

$$\frac{d\eta(t)}{dt} = \theta_1 + 2\theta_2 t = 0 \tag{6.4}$$

thus $t^* = -\frac{\theta_1}{2\theta_2}$ is the location of the maximum in the search direction. In practice, t^* cannot be obtained using equation (6.4) because the parameters θ_0, θ_1 and θ_2 are unknown. The RPR addresses this problem by:

1. Using the observations $Y(t)$ taken along the steepest ascent search to recursively fit the model in (6.3).

2. Testing whether or not the derivative in (6.4) is negative, i.e. test $H_0 : \theta_1 + 2\theta_2 t \geq 0$ vs. $H_a : \theta_1 + 2\theta_2 t < 0$.

For fitting model (6.3), the rule proposes to obtain an estimate of θ_0 from the arithmetic mean of the center points of the previous factorial, denoted $\widehat{Y}(\mathbf{0})$. An estimate of θ_1 is also needed. This is the slope of the response function at the origin in the direction of steepest ascent. Therefore, an estimate of θ_1 is given by the norm of the gradient from the previously fitted first order model:

$$\widehat{\theta}_1 = \|\mathbf{b}\| = \sqrt{b_1^2 + b_2^2 + \cdots + b_k^2} \ .$$

Since in general the previous factorial experiment does not supply any information regarding the second order coefficient θ_2, the RPR uses a Recursive Least Squares (**RLS**) algorithm [4,5,6] to re-estimate the parameter θ_2 at each step t of the steepest ascent search. This allows the RPR to model curvature present in the true response $\eta(t)$. Since the estimate of θ_2 changes after each step, it will be denoted by $\widehat{\theta}_2^{(t)}$. The RLS method requires updates

on the scaled variance of this estimate, for which we will use the notation $P_t \equiv \frac{1}{\sigma_\epsilon^2} Var\left(\widehat{\theta}_2^{(t)}\right)$.

To start the RLS algorithm, an estimate of θ_2 at $t = 0$ (the origin of the search) $\widehat{\theta}_2^{(0)}$, and its associated scaled variance, P_0, are needed. It was suggested to use $\widehat{\theta}_2^{(0)} = -\widehat{\theta}_1/2t_{prior}$, where t_{prior} is an initial guess of how many steps away the directional optimum is located. Although this guess needs to be specified by the user, its impact on the overall performance of the recursive parabolic rule was reported in [37] to be almost insignificant and much less relevant than the effect of κ on the Myers-Khuri rule. The values of P_t decrease so rapidly that the value given to P_0 has practically no effect on the performance of the method [37].

The detection of an eventual drop in the response is based on testing if the first derivative of the fitted model given in equation (6.4) is negative.

From the preceding discussion, the recursive parabolic rule can be formally stated as:

1. Perform an experiment at step t.

2. With the observed value $Y(t)$, update the estimate of $\widehat{\theta}_2^{(t)}$ as follows:

$$\widehat{\theta}_2^{(t)} = \widehat{\theta}_2^{(t-1)} + \frac{P_{t-1}t^2}{1 + t^4 P_{t-1}}(Y(t) - \widehat{Y}(0) - \widehat{\theta}_1 t - \widehat{\theta}_2^{(t-1)}t^2) \; . \qquad (6.5)$$

3. Update P_t, the scaled variance of $\theta_2^{(t)}$ after t experiments:

$$P_t = \left(1 - \frac{P_{t-1}t^4}{1 + t^4 P_{t-1}}\right) P_{t-1} \; . \qquad (6.6)$$

4. If:

$$\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)} t < -3\sqrt{\widehat{\sigma}^2_{\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)} t}} \; , \qquad (6.7)$$

stop the search and return $\widehat{t}^*$ such that $Y(\widehat{t}^*) = max_{\,l=1,..,t}\{Y(l)\}$. Otherwise, increase the step counter (t) and go back to 1. Here, $\widehat{\sigma}^2_{\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)} t}$ denotes the variance of $\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)} t$.

Steps 2 and 3 constitute a standard Recursive Least Squares update for the case of one parameter, θ_2. In this rule, θ_1 and $\widehat{Y}(0)$ are not recursively updated at each step.

Step 4 in the algorithm considers the sampling variability of $\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)}t$. Assuming θ_1 "known" and equal to $\widehat{\theta}_1$, we have that $Var(\widehat{\theta}_1) = 0$ and:

$$\widehat{\sigma}^2_{(\widehat{\theta}_1+2\widehat{\theta}_2^{(t)}t)} = \frac{120t\widehat{\sigma}^2_\epsilon}{(t+1)(2t+1)(3t^2+3t-1)} \approx 4\widehat{\sigma}^2_\epsilon t^2 P_t \qquad (6.8)$$

The estimate of the random noise variance, $\widehat{\sigma}^2_\epsilon$, is obtained from the residual variance of the previous factorial. Assuming this variance also known and equal to $\widehat{\sigma}^2_\epsilon$, step 4 becomes a size $\alpha = 0.0013$ test of the null hypothesis $H_0 : \theta_1 + 2\theta_2 t \geq 0$ vs. the alternative $H_a : \theta_1 + 2\theta_2 t < 0$. When H_0 is rejected we conclude that a drop in the mean response was observed and the search is stopped.

The main advantage of the recursive parabolic rule is that it makes use of all the information available from the previous factorial experiment that was conducted, which is used to estimate θ_0, θ_1 and σ^2_ϵ. Only θ_2 is estimated during the search. This type of efficiency can be very valuable when the cost of an experiment is high. However, the performance of the RPR (6.5–6.7) deteriorates under non-quadratic responses. We illustrate this rule next for a non-quadratic response. For additional examples and derivations see [37].

Example. Use of the Recursive Parabolic Rule. Suppose the true response function consists of a quartic polynomial on two controllable factors. Figure 6.2 contains a surface plot of this function. The true maximum is located at $\mathbf{x}^* = [7.87\,,\ 12.89]$. A random noise term, $\epsilon \sim N(0, 30)$, was added to the true response function to simulate the observed values.

A 2-level full factorial with 2 center points was performed at the location marked with the "**o**" in Figure 6.2. The straight line represents the steepest ascent direction, $\mathbf{b} = [-14.1,\ 23.1]$, obtained from the first order model. Also from this first order model we obtain: $\widehat{\theta}_1 = 27.04$, $\widehat{Y}(\mathbf{0}) = -36.18$ and $\widehat{\sigma}_\epsilon = 5.98$. In addition we use $P_0 = 10$. Suppose in this example that the true response in the steepest ascent direction obeys:

$$\eta(t) = -36.4 + 25.9t - 3.69t^2 + 0.2903t^3 - 0.0083t^4,$$

which has a maximum at $t^* = 15.1$ and corresponds to a mean response $\eta(t^*) = 83.86$. However, let us suppose, for illustration purposes, that we choose $t_{prior} = 18$. This implies that $\widehat{\theta}_2^{(0)} = -\widehat{\theta}_1/2t_{prior} = -0.75$.

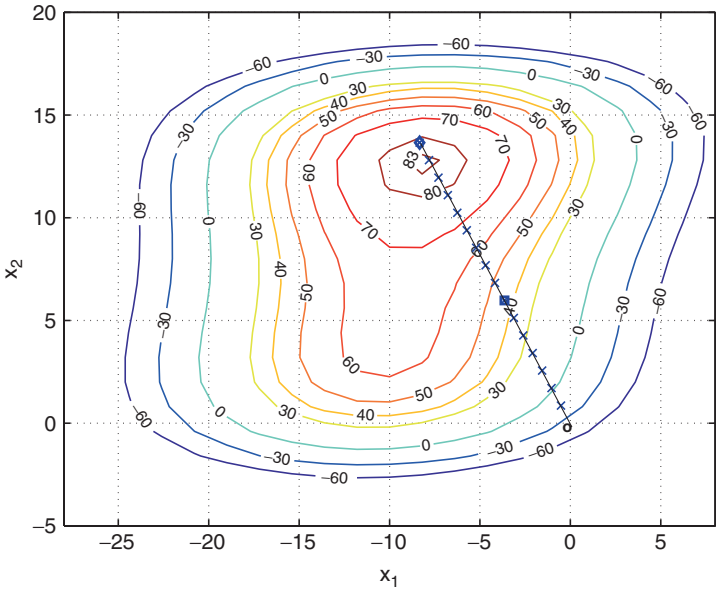


Figure 6.2. Surface plot of the true response response function for the example. The "**o**" marks the center of the initial factorial design, the x's mark the experiments conducted along the steepest ascent search, the square marks the stopping point of the RPR rule and the diamond marks the stopping point of the ERPR (from [102])

Table 6.1 shows the computations required by the RPR at each step t and Figure 6.3 shows the realizations of $Y(t)$ and $\widehat{\theta}_1 + 2\widehat{\theta}_2^{(t)} t$. It is observed that the RPR rule stops at $t_{stop} = 7$ and returns $\widehat{t}^* = 6$ as the best point (step) along the steepest ascent direction. Both of these values are considerable different than the true optimum $t^* = 15.1$. The expected value of the response at $\widehat{t}^*$, $\eta(\widehat{t}^*) = 38.17$, is significantly smaller than the true optimum value, $\eta(t^*) = 83.86$. The poor performance of the RPR is due to the non-quadratic behavior of the true response. ■

An Enhanced Recursive Parabolic Rule (ERPR)

Response Surface Methods assume the validity of second order models only locally. However, during steepest ascent searches changes over large ranges on the controllable factor space can be made to the factors, contradicting the local model assumption. Therefore, for relatively long searches, and specially when the noise standard deviation, σ_ϵ, is small, the RPR can become sensitive to

Table 6.1. Computations for the application of the Recursive Parabolic Rule in the example (from [102])

t	$Y(t)$	$\widehat{\theta}_2^{(t)}$	P_t	$\widehat{\theta}_1+2\widehat{\theta}_2^{(t)}t$	$\widehat{\sigma}^2_{\widehat{\theta}_1+2\widehat{\theta}_2^{(t)}t}$	$-3\sqrt{\widehat{\sigma}^2_{\widehat{\theta}_1+2\widehat{\theta}_2^{(t)}t}}$
0	-36.18	-0.75	$10.0 \times 10^{+1}$	27.04		
1	-6.13	2.67	9.09×10^{-1}	32.38	143.04	-35.88
2	-0.46	-4.12	5.85×10^{-2}	10.55	33.66	-17.40
3	32.95	-1.82	1.02×10^{-2}	16.13	13.14	-10.87
4	21.45	-2.79	2.82×10^{-3}	4.74	6.47	-7.63
5	38.99	-2.54	1.02×10^{-3}	1.63	3.65	-5.73
6	47.69	-2.33	4.40×10^{-4}	-0.96	2.26	-4.51
7	41.12	-2.31	2.14×10^{-4}	-5.28	1.50	-3.67

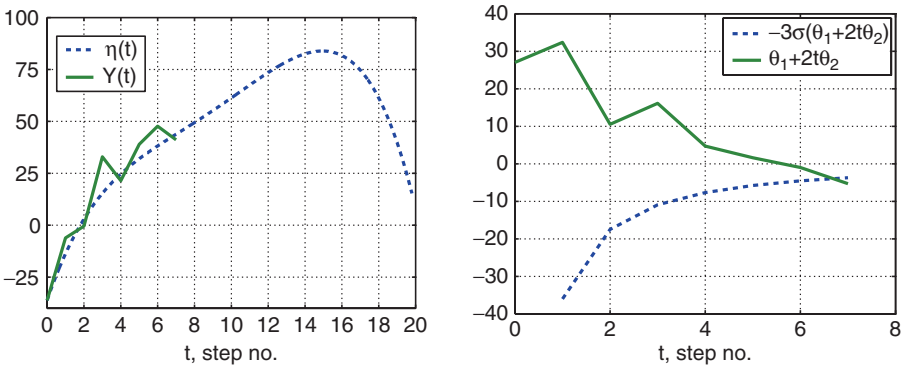


Figure 6.3. Illustration of the RPR. On the left, realization of $Y(t)$ and values of $\eta(t)$. On the right, the estimated first derivative and the stopping limits (dotted line) computed using (6.7). Source: [102]

non-quadratic behavior which can lead to stopping before the maximum. This situation was observed in the example in the previous section.

An enhanced recursive parabolic rule (ERPR) that is more robust to non-quadratic behavior was proposed by Miro et al. [102] by making the following modifications to the recursive rule in [37]:

1 Recursively fit the intercept and the first order term in addition to the second order term in equation (6.3). This should increase the robustness against non-quadratic behavior.

2 Allow the specification of a maximum number of experiments in the Recursive Least Squares algorithm, defining what from now on will be called a "window", in order to fit only a *local* parabolic model along the search direction. This model should be less sensitive to large scale deviations from quadratic behavior.

3 Use a coding scheme on the step numbers in order to obtain a near orthogonal design, reducing in this form the bias and variance of the parameter estimates of model (6.3).

While at first sight this seems to result in an overly complicated approach, particularly compared to traditional steepest ascent practice, it was shown in [102] how the computations simplify notably and are therefore implementable without any specialized software. Each of these 3 modifications are described and illustrated next. Readers who want to skip the technical details may wish to read the "Summary of computations" in Section 6.1.3, shown later in this chapter, and study only the examples in what follows next.

Modification 1: Using Recursive Least Squares to update the estimates of θ_0, θ_1 and θ_2.

In general, when the true function is no longer quadratic, fitting all the parameters in a parabolic model will give a better local approximation of the true function, since the model has more flexibility to adapt to the changes in curvature of a non-quadratic function.

The RLS algorithm for updating the three parameters θ_0, θ_1 and θ_2, makes use of the following definitions:

$$\hat{\boldsymbol{\theta}}^{(t)} = \begin{bmatrix} \widehat{\theta}_0^{(t)} \\ \widehat{\theta}_1^{(t)} \\ \widehat{\theta}_2^{(t)} \end{bmatrix}, \boldsymbol{\phi}_t = \begin{bmatrix} 1 \\ t \\ t^2 \end{bmatrix}, \frac{d\boldsymbol{\phi}_t}{dt} \equiv \mathbf{d}_t = \begin{bmatrix} 0 \\ 1 \\ 2t \end{bmatrix}$$

$$\text{and } \mathbf{P}_t = Var\left(\hat{\boldsymbol{\theta}}^{(t)}\right)/\sigma_\epsilon^2$$

Then $\frac{dY(t)}{dt} = \mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)}$, and we can write the recursive parabolic rule for all three parameters θ_0, θ_1 and θ_2 in the following way:

1. Perform an experiment at step t.

2. With the observed value $Y(t)$, update the estimate of $\widehat{\boldsymbol{\theta}}(t)$ as follows:

$$\hat{\boldsymbol{\theta}}^{(t)} = \hat{\boldsymbol{\theta}}^{(t-1)} + \frac{\mathbf{P}_{t-1}\boldsymbol{\phi}_t}{1 + \boldsymbol{\phi}_t'\mathbf{P}_{t-1}\boldsymbol{\phi}_t}\left(Y(t) - \boldsymbol{\phi}_t'\hat{\boldsymbol{\theta}}^{(t-1)}\right) . \tag{6.9}$$

3. Update $\mathbf{P}_t$, the scaled covariance matrix of $\boldsymbol{\theta}^{(t)}$ after t experiments:

$$\mathbf{P}_t = Var\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)/\sigma_\epsilon^2 = \left(\mathbf{I} - \frac{\mathbf{P}_{t-1}\boldsymbol{\phi}_t}{1 + \boldsymbol{\phi}_t'\mathbf{P}_{t-1}\boldsymbol{\phi}_t}\boldsymbol{\phi}_t'\right)\mathbf{P}_{t-1} . \tag{6.10}$$

4. If:

$$\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)} < -1.645\hat{\sigma}_\epsilon\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t} \ , \tag{6.11}$$

stop the search and return $\widehat{t^*}$ such that $Y(\widehat{t^*}) = max_{l=1,..,t}\{Y(l)\}$. Otherwise, increase the step counter (t) and go back to 1.

The -1.645 multiplier in equation (6.11) implies a size $\alpha = 0.05$ test. This is done because it is not assumed that θ_0 and θ_1 are "known" as it is assumed in the RPR. Therefore, there is no need to protect against underestimating the variance of the first derivative's estimate. It was observed in simulation studies that this selection for the size of the test makes the rule more effective in short searches. Computations (6.9–6.11) are very easy to implement on a spreadsheet software.

For the initial values of the parameter estimates $\widehat{\theta}_0^{(0)}$, $\widehat{\theta}_1^{(0)}$ and $\widehat{\theta}_2^{(0)}$, a similar approach than in the RPR rule was suggested:

$$\widehat{\theta}_0^{(0)} = \widehat{Y}(\mathbf{0}) \tag{6.12}$$

$$\widehat{\theta}_1^{(0)} = \|\mathbf{b}\| \tag{6.13}$$

$$\widehat{\theta}_2^{(0)} = -\frac{\widehat{\theta}_1^{(0)}}{2t_{prior}} \tag{6.14}$$

For the initial scaled covariance matrix of the parameter estimates, $\mathbf{P}_0$, Miro et al. [102] suggested to use:

$$\mathbf{P}_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10 \end{bmatrix} \tag{6.15}$$

The large value of 10 given to $Var\left(\widehat{\theta}_2^{(0)}\right)/\sigma_\epsilon^2$ makes the rule robust against possibly large discrepancies between t_{prior} and t^*, providing some "adaptation"

ability to varying curvature. The smaller scaled initial variances of the other two parameters assure that the search will not stop too early before the maximum (see [102]).

Modification 2: Use of a moving window.

Goodwin and Payne [63] suggested to use a moving "window" in recursive least squares estimation in order to track time-changing parameters in adaptive control applications. The name is taken from the fact that *only a fixed number of observations –a window–is used in the RLS algorithm*. The effect of the moving window is that the algorithm discards the last observation once a new observation has been made, keeping the number of observations used at any step (i.e. the "window" size) constant. This can be used to fit a local model during a steepest ascent search. In such case, the change or *adaptation* in the parameter estimates allows to model possible changes in curvature due to non-quadratic behavior in the controllable factor space.

The size of the window must be selected in order to make a compromise between bias and variance. A large window will use much of the response information available giving a very small variance for the first derivative and hence a very powerful test for a negative derivative, but it will provide biased estimates for curvature.

One way of making a compromise between bias and variance, when the form of $\eta(t)$ is unknown, is to determine the minimum number of observations N that gives a pre-specified power for the test of a negative derivative. Miro et al. [102] gave the folowing expression for the power of an $\alpha = 0.05$ test of the hypothesis $H_o : \theta_1 + 2\theta_2 t \geq 0$ versus an alternative $H_a : \theta_1 + 2\theta_2 t < 0$:

$$K(\Delta) = \Pr_{\Delta}\left(\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)} < -1.645\widehat{\sigma}_{\epsilon}\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t}\right) \tag{6.16}$$

where Δ is the expected value of the derivative of the response along the search direction, i.e., $\Delta = E\left(\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)}\right)$. Now, if it turns out that $H_a : \Delta < 0$ is true, we have that:

$$K(\Delta) = \Pr_{\Delta}\left(\frac{\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)} - \Delta}{\widehat{\sigma}_{\epsilon}\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t}} < -1.645 - \frac{\Delta}{\widehat{\sigma}_{\epsilon}\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t}}\right). \tag{6.17}$$

Under the assumption of a locally valid second order model, $E\left(\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)}\right) = \mathbf{d}_t'\boldsymbol{\theta}^{(t)}$. In addition, as in the RPR, assuming σ_ϵ^2 "known" and equal to $\widehat{\sigma}_\epsilon^2$ we have that:

$$\frac{\mathbf{d}_t'\hat{\boldsymbol{\theta}}^{(t)} - \Delta}{\widehat{\sigma}_\epsilon\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t}} = Z \sim N(0,1) \ . \tag{6.18}$$

Given that

$$\mathbf{d}_t'\mathbf{P}_t\mathbf{d}_t \approx \frac{12(2t-1)(8t-11)}{(t-1)(t-2)(t+2)(t+1)t} \tag{6.19}$$

and substituting (6.19) in (6.17) we have that:

$$1-\beta = \Phi\left(-1.645 - \frac{\Delta}{\widehat{\sigma}_\epsilon}\sqrt{\frac{(t-1)(t-2)(t+2)(t+1)t}{12(2t-1)(8t-11)}}\right) \tag{6.20}$$

where β is the probability of a Type II error, $1-\beta$ is the power of the test, and Φ is the standard normal CDF. The window size N is then set to the value of t that solves equation (6.20) for a desired value of power. Provided N is not too large, the bias due to higher order effects will be significantly reduced when compared to the case when all the observed values $Y(t)$ are used.

In practice, it will be difficult to come up with suitable values of Δ. Miro et al. [102] recommended that the value of Δ be specified as a percentage of the norm of the gradient at the beginning of the search, that is:

$$\Delta = -a\|\mathbf{b}\|$$

where a is a number between 0 and 1 decided by the experimenter. This choice guarantees with some probability the rejection of the null hypothesis when the response is dropping at a $100a\%$ of the rate at which it was increasing at the beginning of the search.

Example. Updating all parameter estimates and use of a moving window. We now show an example of how to apply the first two modifications to the RPR rule with the same example used earlier in the previous section. For determining the window size N we chose $a = 0.4$ and $1-\beta = 0.8$ (this selection was justified based on an extensive simulation study conducted by G. Miro [101]). This means we want a window size such that it allows to detect with power

equal to 0.80 a response that is dropping 40% of what was increasing initially. With these selections we have that $\Delta = -a\|\mathbf{b}\| = -10.82$ and equation (6.20) becomes:

$$0.8 = \Phi\left(-1.645 - \frac{-10.82}{5.98}\sqrt{\frac{(t-1)(t-2)(t+2)(t+1)t}{12(2t-1)(8t-11)}}\right)$$

The value $t = 6.7070$ solves the previous equation, and therefore we set a window size of $N = 7$. Therefore, we use the recursive computations for $t < N - 1 = 6$, or $t \leq 5$. Table 6.2 contains the required computations obtained. As it can be seen, the proposed rule does not signal to stop the search for $t \leq 5$. ■

Modification 3: Use of a coding convention.

The use of a coding convention is frequently recommended in response surface methods. A coding convention reduces the bias in the parameter estimates [8,9]. The single regressor t (step number along the search) used in the recursive parabolic rule can be coded using the sequence:

$$\left\{-\frac{t-1}{2}, -\frac{t-3}{2}, -\frac{t-5}{2}, \ldots, \frac{t-5}{2}, \frac{t-3}{2}, \frac{t-1}{2}\right\} \tag{6.21}$$

This is a sequence of t numbers centered at zero that provides a new scale. It was shown in [102] how using this coding convention helps to reduce the bias in the estimate of the first derivative due to the presence of third and fourth order effects, i.e. when the true response is not quadratic.

An added advantage of this coding convention for t is that if used for $t \geq N - 1$ then the modified recursive procedure (equations (6.9) to (6.11)) simplifies significantly. Miro [101] realized that by using this coding, the design matrix $\mathbf{X}$ remains *fixed* at:

$$\mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} & t & t^2 \end{array} \\ \begin{bmatrix} 1 & \frac{(N-1)}{2} & \frac{(N-1)^2}{4} \\ 1 & \frac{(N-3)}{2} & \frac{(N-3)^2}{4} \\ \vdots & \vdots & \vdots \\ 1 & -\frac{(N-3)}{2} & \frac{(N-3)^2}{4} \\ 1 & -\frac{(N-1)}{2} & \frac{(N-1)^2}{4} \end{bmatrix} \end{array}$$

Table 6.2. Computations for the enhanced rule (ERPR) for steps $t < N - 1$. The last column gives the stopping limits (from [102])

t	$Y(t)$	$\widehat{\theta}_0^{(t)}$	$\widehat{\theta}_1^{(t)}$	$\widehat{\theta}_2^{(t)}$	$\mathbf{P}_t$	$\mathbf{d}_t'\widehat{\boldsymbol{\theta}}^{(t)}$	$\widehat{\sigma}_\epsilon^2\mathbf{d}_t'\mathbf{P}_t\mathbf{d}$	$-1.645\widehat{\sigma}_\epsilon\sqrt{\mathbf{d}_t'\mathbf{P}_t\mathbf{d}}$
0.00	−36.18	−36.18	27.04	−0.75	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10 \end{bmatrix}$	27.04		
1.00	−6.13	−35.89	27.33	2.14	$\begin{bmatrix} 0.92 & -0.08 & -0.77 \\ -0.08 & 0.92 & -0.77 \\ -0.77 & -0.77 & 2.31 \end{bmatrix}$	31.62	254.8	−26.26
2.00	−0.46	−33.19	28.86	−5.95	$\begin{bmatrix} 0.70 & -0.20 & -0.10 \\ -0.20 & 0.85 & -0.39 \\ -0.10 & -0.39 & 0.29 \end{bmatrix}$	5.04	86.6	−15.31
3.00	32.95	−36.09	24.64	−0.95	$\begin{bmatrix} 0.63 & -0.30 & 0.02 \\ -0.30 & 0.71 & -0.22 \\ 0.02 & -0.22 & 0.09 \end{bmatrix}$	18.93	44.0	−10.91
4.00	21.45	−34.99	28.88	−3.41	$\begin{bmatrix} 0.62 & -0.34 & 0.04 \\ -0.34 & 0.55 & -0.13 \\ 0.04 & -0.13 & 0.03 \end{bmatrix}$	1.61	27.2	−8.58
5.00	38.99	−34.94	26.29	−2.44	$\begin{bmatrix} 0.62 & -0.34 & 0.04 \\ -0.34 & 0.42 & -0.08 \\ 0.04 & -0.08 & 0.02 \end{bmatrix}$	1.91	18.3	−7.04

The scaled covariance matrix of the parameters estimates *also remains constant* and is given by:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{3(3N^2-7)}{4N(N^2-4)} & 0 & -\frac{15}{N(N^2-4)} \\ 0 & \frac{12}{N(N^2-1)} & 0 \\ -\frac{15}{N(N^2-4)} & 0 & \frac{180}{N(N^4-5N^2+4)} \end{bmatrix}$$

In general, there will be small differences between $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{P}_t$ for $t \geq N-1$ due to the effect of the initial value $\mathbf{P}_0$. Using the coding convention (6.21) with a window size N, step $(N-1)/2$ corresponds to the end of the window at which point the vector of the derivatives of the regressors equals:

$$\mathbf{d}_N = \begin{bmatrix} 0 \\ 1 \\ (N-1) \end{bmatrix}.$$

Defining $\mathbf{b}_N \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{d}_N$, we can obtain updates for the value of the first derivative of the response by:

$$\mathbf{d}_N'\widehat{\boldsymbol{\theta}}^{(t)} = \mathbf{d}_N'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_N(t) = \mathbf{b}_N'\mathbf{Y}_N(t)$$

where the "moving window" of the last N observations is given by:

$$\mathbf{Y}_N(t) = \begin{bmatrix} Y(t) \\ Y(t-1) \\ \vdots \\ Y(t-N+1) \end{bmatrix}.$$

To obtain an update of the first derivative of the response for steps $t \geq N-1$, the user only has to compute the inner product between $\mathbf{b}_N$ and $\mathbf{Y}_N(t)$. Furthermore, for $t \geq N-1$ the scaled variance of the first derivative, denoted by v_N, *remains fixed* at:

$$v_N \equiv \mathbf{d}_N'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{d}_N = \frac{12(2N-1)(8N-11)}{(N-1)(N-2)(N+1)(N+1)N},$$

(note how it is not a function of t). Therefore, for $t \geq N-1$ the enhanced recursive rule is very easy to use since $\mathbf{b}_N$ and v_N can be determined *before starting the search* once the window size N is determined. Table 6.3 provides values for $\mathbf{b}_N$ and v_N. The values of N were calculated using $a = 0.4$ and

Table 6.3. Values of window size N, the $N \times 1$ vector $\mathbf{b}_N$ and the scalar $\sqrt{v_N}$ needed to utilize equation (6.22) in the enhanced stopping rule (ERPR). N was calculated assuming $a = 0.4$ and $1 - \beta = 0.8$. Source: [102]

		Signal-to-Noise Ratio $\|\mathbf{b}\|/\widehat{\sigma}_\epsilon$.												
Lower Bound		10.22	8.12	6.02	4.72	3.84	3.20	2.73	2.36	2.07	1.84	1.64	1.48	
Upper Bound			10.21	8.11	6.01	4.71	3.83	3.19	2.72	2.35	2.06	1.83	1.63	1.47
Window Size (N)		3	4	5	6	7	8	9	10	11	12	13	14	15
$\sqrt{v_N}$		2.550	1.565	1.115	0.853	0.681	0.562	0.474	0.407	0.354	0.313	0.278	0.250	0.226
	$Y(t)$	0.50	1.05	0.77	0.59	0.46	0.38	0.31	0.26	0.22	0.19	0.16	0.14	0.13
	$Y(t-1)$	−2.00	−0.65	−0.19	0.00	0.07	0.10	0.11	0.11	0.11	0.10	0.09	0.09	0.08
	$Y(t-2)$	1.50	−0.85	−0.57	−0.33	−0.18	−0.09	−0.04	0.00	0.02	0.03	0.03	0.04	0.04
	$Y(t-3)$		0.45	−0.39	−0.39	−0.29	−0.20	−0.13	−0.08	−0.05	−0.03	−0.01	0.00	0.01
	$Y(t-4)$			0.37	−0.18	−0.25	−0.22	−0.17	−0.13	−0.10	−0.07	−0.05	−0.03	−0.02
	$Y(t-5)$				0.30	−0.07	−0.16	−0.16	−0.14	−0.12	−0.09	−0.07	−0.06	−0.04
$\mathbf{b}_N$	$Y(t-6)$					0.25	−0.02	−0.10	−0.12	−0.11	−0.10	−0.08	−0.07	−0.06
	$Y(t-7)$						0.21	0.01	−0.06	−0.09	−0.09	−0.08	−0.07	−0.06
	$Y(t-8)$							0.18	0.03	−0.04	−0.06	−0.07	−0.07	−0.06
	$Y(t-9)$								0.15	0.03	−0.02	−0.05	−0.06	−0.06
	$Y(t-10)$									0.13	0.04	−0.01	−0.03	−0.04
	$Y(t-12)$										0.11	0.04	0.00	−0.02
	$Y(t-13)$											0.10	0.04	0.00
	$Y(t-14)$												0.09	0.04
	$Y(t-15)$													0.08

$1-\beta = 0.8$ as before. To use this table, one needs to find the pair of lower and upper bounds that brackets the signal-to-noise ratio $\|\mathbf{b}\|/\widehat{\sigma}_\epsilon$ obtained from the previous factorial. From that column, the window size N, the scaled standard deviation of the first derivative, $\sqrt{v_N}$, and the $N \times 1$ vector b_N are read. For the previous example, we have that $\|\mathbf{b}\|/\widehat{\sigma}_\epsilon = 4.52$, which lies between 3.84 and 4.71, hence $N = 7$.

For $t \geq N$ the enhanced recursive parabolic rule is as follows:

1. Perform an experiment at step $t \geq N-1$

2. With the observed value $Y(t)$, update vector $\mathbf{Y}_N(t)$ by discarding its first, shifting the remaining elements one position up in the vector and including $Y(t)$ as the last element in $\mathbf{Y}_N(t)$.

3. Read $\mathbf{b}_N$ and v_N from Table 6.3 and test if:

$$\mathbf{b}_N' \mathbf{Y}_N(t) < -1.645 \widehat{\sigma}_\epsilon \sqrt{v_N} \quad , \tag{6.22}$$

If the inequality holds, then stop the search and return $\widehat{t}^*$ such that $Y(\widehat{t}^*) = max_{\ l=1,..,t}\{Y(l)\}$. Otherwise, increase the step counter (t) and go back to 1.

Example. Enhanced recursive parabolic rule for $t \geq N-1$. To illustrate, let us continue the previous example for steps $t \geq N-1 = 6$. From Table 6.3 we read $\mathbf{b}_7 = [0.46, 0.07, -0.18, -0.29, -0.25, -0.07, 0.25]'$ and $\sqrt{v_7} = 0.681$. From this, the limit in (6.22) equals to -6.69 (recall that $\widehat{\sigma}_\epsilon = 5.98$ from the previous examples).

Table 6.4 contains the computations necessary to implement the rule. The values of $\widehat{\theta}_0(t)$, $\widehat{\theta}_1(t)$ and $\widehat{\theta}_2(t)$ are not needed in the proposed methodology for $t \geq N-1$ so they are not shown. However, the procedure is implicitly re-estimating these parameters, adapting to the non-constant curvature of the underlying quartic model. Figure 6.4 illustrates the realizations of $Y(t)$ and $\mathbf{d}_t' \boldsymbol{\theta}^{(t)}$.

The proposed rule stops at $t_{stop} = 16$ and returns $\widehat{t}* = 14$, only 1.1 steps away from $\widehat{t}* = 15.1$. The expected mean at $\widehat{t}*$, is $\eta(\widehat{t}*) = 82.54$, only 1.5% smaller than at $\widehat{t}*$. In addition, notice that the rule stops after an increase in the observed response. ■

Table 6.4. Computations for the enhanced rule (ERPR) for steps $t \geq 6$. The limit is $-1.645\widehat{\sigma}_\epsilon\sqrt{v_7} = -6.69$ for all $t \geq 6$. Third column shows the "rectangular window" of the last $N = 7$ observations. Source: [102]

t	$Y(t)$	$\mathbf{Y_7(t)}$	$\mathbf{b_7'Y_7(t)}$
6	47.69	[−36.18, −6.13, −0.46, 32.95, 21.45, 38.99, 47.69]	3.19
7	41.12	[−6.13, −0.46, 32.95, 21.45, 38.99, 47.69, 41.12]	−0.33
8	47.42	[−0.46, 32.95, 21.45, 38.99, 47.69, 41.12, 47.42]	−2.53
9	52.04	[32.95, 21.45, 38.99, 47.69, 41.12, 47.42, 52.04]	3.54
10	58.59	[21.45, 38.99, 47.69, 41.12, 47.42, 52.04, 58.59]	1.36
11	75.08	[38.99, 47.69, 41.12, 47.42, 52.04, 58.59, 75.08]	12.26
12	76.40	[47.69, 41.12, 47.42, 52.04, 58.59, 75.08, 76.40]	12.63
13	97.27	[41.12, 47.42, 52.04, 58.59, 75.08, 76.40, 97.27]	14.35
14	100.80	[47.42, 52.04, 58.59, 75.08, 76.40, 97.27, 100.80]	12.14
15	82.84	[52.04, 58.59, 75.08, 76.40, 97.27, 100.80, 82.84]	−3.48
16	82.94	[58.59, 75.08, 76.40, 97.27, 100.80, 82.84, 82.94]	−11.18

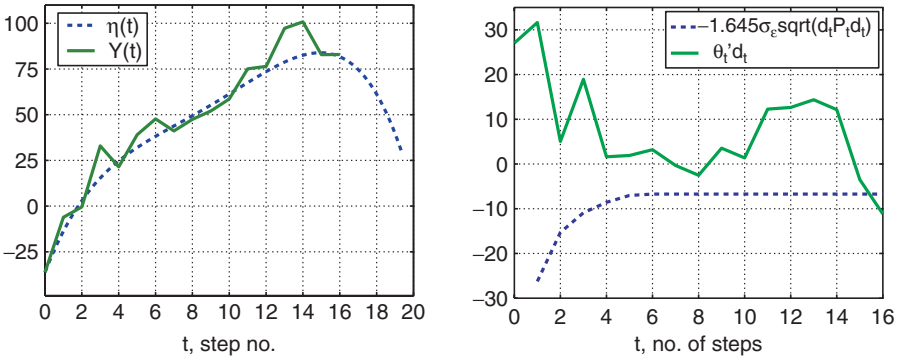


Figure 6.4. Graphical illustration of the enhanced recursive rule (ERPR). On the left, realization of $Y(t)$ and values of $\eta(t)$. On the right, illustration of stopping rule (6.11) for $t < 6$ and stopping rule (6.22) for $t \geq 6$. Graph on the right shows the estimated derivative and the stopping limits (dotted line). Note how the limit is a constant for $t \geq 6 = N - 1$. Source: [102]

6.1.3 Summary of Computations, ERPR Method

1. Perform a 2-level factorial design with replicated center points at the baseline operating conditions (point $t = 0$). Compute $||\boldsymbol{b}|| = ||\widehat{\boldsymbol{\beta}}|| = \sqrt{\sum_{i=1}^{k} b_i^2}$. Estimate $\widehat{\sigma}_\varepsilon^2$ from the standard deviation of the center points.

2. Compute $||\boldsymbol{b}||/\widehat{\sigma}_\varepsilon$ and use Table 6.3 to find N, the window size, the $\boldsymbol{b}_N$ vector, and the v_N upper bound.

3. Initialize the recursive computations ($t = 0$). Define t_{prior}. Use equations (6.12) to (6.15).

4. For $t = 1$ to $t = N - 1$, perform an experiment at step t along the ascent direction. Use the recursive computations (6.9) and (6.10) to update the parameter estimates and covariance matrix. If (6.11) is true, make $t_{stop} = t$ and go to 8.

5. Load vector $\mathbf{Y}_N(t)$ with observations $\overline{Y}(0), Y(1), \ldots, Y(N-1)$.

6. For $t \geq N - 1$, perform an experiment at step t along the ascent direction and update vector $\mathbf{Y}_N(t)$ using a moving window of size N (i.e., discard its first element, shift all elements up one position and add $Y(t)$ at its end).

7. Test for termination at each $t \geq N - 1$: if

$$\boldsymbol{b}_N' \mathbf{Y}_N(t) < -1.645 \widehat{\sigma}_\varepsilon \sqrt{v_N},$$

stop ($t_{stop} = t$).

8. Return $\widehat{t}^*$ such that $Y(\widehat{t}^*) = \max_{l=1,2,\ldots,t_{stop}} Y(l)$

Step 1 is common to basic steepest ascent. Special computations are required in steps 4, 6, and 7. They are very simple to perform and can be easily implemented in a spreadsheet software (see Problem 17).

As mentioned earlier, Miro [101] conducted extensive simulations of the ERPR, MK, and RPR stopping schemes. His results indicate that the ERPR stopping scheme provides best performance, including most robust behavior against non-normality (a problem for the MK rule) and against non-quadratic behavior (a problem for the RPR rule).

6.2 Confidence Cone for the Direction of Steepest Ascent/Descent

Selecting the step size in a steepest ascent/descent search ultimately depends on how much confidence do we have in such direction. If the model fit is poor, the parameter estimates for the main effects, and hence, the gradient estimate will be poor (large standard error), and this means that we should not trust the estimated steepest ascent/descent direction very much. Evidently, if the fit is *very* poor it is either because we need to conduct more experiments to reduce

the noise and get a better fit of the linear model or because the linear model (and steepest ascent/descent) was not adequate in the first place.

There will still be cases in which a first order model fits reasonable well, it shows no lack of fit, yet the fit might not be that great. Compared with a model with really good fit, the first scenario should call for more cautious steps be taken. To answer these questions, Box [17] proposed to compute a confidence "cone" around the direction of steepest ascent/descent. It turns out that any operating conditions $\boldsymbol{x}' = (x_1, x_2, \ldots, x_k)$ that satisfy

$$\sum_{i=1}^{k} b_i^2 - \frac{(\sum_{i=1}^{k} b_i x_i)^2}{\sum_{i=1}^{k} x_i^2} \leq (k-1) s_b^2 F_{\alpha,k-1,v} \tag{6.23}$$

generate a direction that lies inside the $100(1-\alpha)\%$ confidence cone of steepest *ascent* if in addition

$$\sum_{i=1}^{k} b_i x_i > 0,$$

and generate a direction that lies inside the $100(1-\alpha)\%$ confidence cone of steepest *descent* if

$$\sum_{i=1}^{k} b_i x_i < 0.$$

Here, $s_b^2 = \widehat{\text{Var}}(b_i)$ and has v degrees of freedom. It is assumed that the DOE run to fit the model estimates all the main effects with the same precision.

The rationale for result (6.23) is based on Figure 6.5, which shows the vectors involved in the cone expression in some k-dimensional space. The projection of vector $\boldsymbol{b}$ on to $\boldsymbol{x}$ is given by $\frac{\boldsymbol{b}'\boldsymbol{x}}{\|\boldsymbol{x}\|}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and has magnitude $\frac{\boldsymbol{b}'\boldsymbol{x}}{\|\boldsymbol{x}\|}$. Thus, from Pythagoras theorem, we have that

$$\|\boldsymbol{b}\|^2 - \left(\frac{\boldsymbol{b}'\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)^2 = \|\boldsymbol{d}\|^2.$$

Notice how the left hand side is identical to the left hand side of the inequality in (6.23). The direction generated by $\boldsymbol{x}$ is inside the confidence cone around the estimated gradient (given by vector $\boldsymbol{b}$) if the vector $\boldsymbol{d}$ is small. This vector should account for the statistical variability in our estimates $\boldsymbol{b}$. If all b_i are estimated with the same standard error, the variance of b_i, s_b^2, is a measure of this variability. Now, the left hand side of the last expression is the length of vector $\boldsymbol{d}$, and is a sum of squares. Thus, we can compare this with s_b^2 to

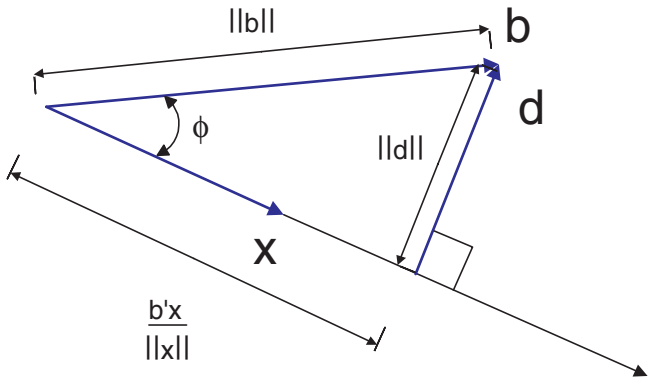


Figure 6.5. The vectors involved in the confidence cone expression

determine if $\boldsymbol{x}$ is inside the cone or not. Since vector $\boldsymbol{d}$ has $(k-1)$ degrees of freedom (its tip must coincide with vector $\boldsymbol{b}$) we should scale the sum of squares that represents its length by $(k-1)$ to form a mean square which can be compared with s_b^2. We then have that, assuming normally distributed errors,

$$\frac{\left(\|\boldsymbol{b}\|^2 - \left(\frac{\boldsymbol{b}'\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)^2\right)/(k-1)}{s_b^2} \sim F_{k-1}, v$$

and the result (6.23) follows.

Since $\boldsymbol{b}'\boldsymbol{x} = (cos\ \phi)\|\boldsymbol{b}\|\|\boldsymbol{x}\|$, then if $\boldsymbol{b}'\boldsymbol{x} > 0 \ \Rightarrow cos\ \phi > 0 \Rightarrow \phi < 90°$, and $\boldsymbol{x}$ is inside the steepest ascent cone. Likewise, if $\boldsymbol{b}'\boldsymbol{x} < 0 \Rightarrow \phi > 90°$ and $\boldsymbol{x}$ is inside the steepest descent cone.

For vector $\boldsymbol{x}$ to lie inside the confidence cone we should have that

$$\|\boldsymbol{d}\|^2 \leq (k-1)s_b^2 F_{\alpha,k-1,v}$$

so we use the square root of the upper bound as the distance value $\|\boldsymbol{d}\|$ in the cone. With this, it is simple to see that the angle ϕ of the cone (Figure 6.5) is

$$\phi = \arcsin\left(\sqrt{\frac{(k-1)s_b^2 F_{\alpha,k-1,v}}{\sum_{i=1}^k b_i^2}}\right).$$

We would like to know how *wide* the confidence cone on the direction of steepest ascent/descent is. If it is too wide, this means the estimated direction is not very precise, and it would be wise not to take big steps in that direction.

It may be even wiser to get a better estimate of the direction through more experiments before conducting a line search.

A quantity useful to determine the precision of a direction is the fraction of directions excluded by the confidence cone, w. If this fraction is close to 1, it means the cone is quite narrow and therefore precise. To compute w, we use:

$$w = 1 - \frac{S_{cap}}{S_{sphere}}$$

where S_{cap} is the surface area of the spherical cap generated by the confidence cone when it reaches a sphere centered at the apex of the cone (see Figure 6.6), and S_{sphere} is the surface area of the sphere.

For k dimensions, the surface area of a hypersphere of unit radius is

$$S_{sphere} = \frac{2\pi^{k/2}}{\Gamma(k/2)}$$

and the surface area of the spherical cap of a hypersphere of unit radius is

$$S_{cap} = \frac{(k-1)\pi^{(k-1)/2}}{\Gamma(\frac{k+1}{2})} \int_0^{\phi} sin^{k-2}u \; du$$

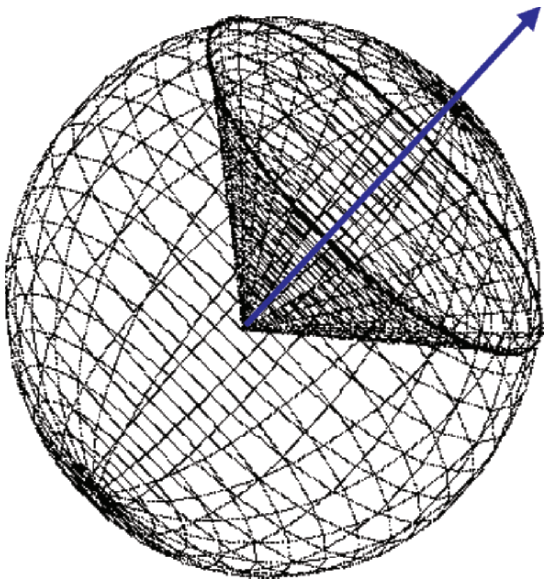


Figure 6.6. The fraction of directions included in the cone equals the area of the spherical cap generated by the cone divided by the total area of the sphere

Table 6.5. The proportion of directions excluded by a confidence cone, w, from the formulae for surface areas of caps and hyperspheres

k	S_{cap}	S_{sphere}	w
2	2ϕ	2π	$1 - \frac{\phi}{\pi}$
3	$2\pi(1 - \cos(\phi))$	4π	$\frac{1}{2}(1 + \cos(\phi))$
4	$2\pi(\phi - \cos(\phi)\sin(\phi))$	$2\pi^2$	$1 + \frac{1}{\pi}(\cos(\phi)\sin(\phi) - \phi)$

Therefore, the fraction of directions excluded by the $100(1-\alpha)\%$ confidence cone of steepest ascent/descent is

$$w = 1 - \frac{(k-1)\Gamma(k/2)}{2\pi^{1/2}\Gamma(\frac{k+1}{2})} \int_0^{\phi} sin^{k-2}u \; du.$$

This expression looks formidable, but for small values of k is quite easy to handle, as shown in Table 6.5.

Based on a characterization of the Student t distribution related to spherical caps, Box and Draper give the equivalent expression:

$$w = 1 - T_{k-1}\left(\sqrt{\frac{\sum b_i^2}{s_b^2 F_{\alpha,k-1,v}} - (k-1)}\right)$$

where $T_{k-1}(u) = P(t_{k-1} \geq u)$ is the complement of a Student t cumulative distribution function with $k-1$ degrees of freedom.

Example. Suppose an experiment with $k = 2$ factors was run, resulting in $s_b^2 = 0.0086$ with $v = n-p = 6$ degrees of freedom, and $b_1 = 0.78$, $b_2 = 0.33$. Thus $\sum b_i^2 = 0.7173$, and using $\alpha = 5\%$, $F_{0.05,1,6} = 5.99$,

$$\phi = \arcsin\left(\sqrt{\frac{(k-1)s_b^2 F_{\alpha,k-1,v}}{\sum_{i=1}^{k} b_i^2}}\right) = 0.2763$$

and since $k = 2$,

$$w = 1 - \frac{\phi}{\pi} = 1 - 0.0879 = 0.9120.$$

Alternatively, using Box and Draper's expression

$$w = 1 - T_1\left[\left(\frac{0.7173}{(0.0086)(5.99)} - 1\right)^{1/2}\right] = 0.9136.$$

The steepest ascent direction eliminates 91% of possible directions, and hence is a relatively precise direction estimator. ■

In practice, if the first order model has a poor fit (a low R^2 value) the corresponding confidence cone will be wide and w small. We mention in passing that the value of w obtained represents a point estimate of the actual percentage of directions excluded from the confidence cone. If experiments are repeated, sampling variability in the parameter estimates will be transmitted to w, hence w will have a sampling distribution which in principle could be studied, although no work has been reported in this sense in the literature. Confidence cones can be useful for steepest ascent/descent when *multiple responses* are considered, see [38]. A procedure for steepest ascent in the presence of noise factors used in a Robust Parameter Design framework has been recently proposed by Miro and del Castillo [103].

In an interesting recent paper, Diaz-Garcia and Ramos Quiroga [46] give the marginal distributions of the coordinates of the maximum point along a steepest ascent search. They also derive the distribution of the maximum response at this point.

6.3 Setting the Step Size and the Direction Simultaneously for New Region Exploration**

Kleijnen et al. [81] have recently proposed[5] an interesting modification of the steepest ascent (SA) procedure that provides a suggestion not only of what step size to use, but also on which direction (other than the SA direction) the search of a new region should be conducted. In this section we present their approach, together with a modification to it by Perez-Salvador and O'Reilly [123]. This section will then be followed by the discussion of a different proposal by Perez-Salvador and O'Reilly [123] on how to choose the step size in SA in order to avoid wild extrapolations of the current model. We assume we wish to *maximize* the response, without loss of generality, and that the factors are coded in the $(-1,1)$ convention.

The idea in [81] is to find the point $\boldsymbol{x}$ where the lower confidence interval on the expected response is a maximum. Such point will give us some assurance (confidence) that the response, on average, at least will not be any lower, since the confidence interval takes into account the variance of the estimated $E[y|\boldsymbol{x}]$.

[5]This section has somewhat more advanced material and may be skipped on a first reading.

Suppose we conduct a first order DOE and fit a linear or first order polynomial model around the current operating conditions. The model is

$$y = \boldsymbol{\beta}'\boldsymbol{x} + \varepsilon, \qquad \boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k),$$

where $\boldsymbol{\beta}$ is a $(k+1) \times 1 = p \times 1$ vector of parameters Write the $p \times 1$ vector of regressors (vector of factors in model form) as $\boldsymbol{x}' = (1, \boldsymbol{d})$ where $\boldsymbol{d}$ is a $k \times 1$ vector. Then, for the assumed model, let $\boldsymbol{X}$ be the design matrix with columns that correspond to the model. Then we can write

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} 1/n & \boldsymbol{b}' \\ \boldsymbol{b} & \boldsymbol{C} \end{bmatrix}$$

where n is the number of experiments in the DOE, $\boldsymbol{b}$ is a $k \times 1$ vector of constants and $\boldsymbol{C}$ is a $k \times k$ matrix. With this notation, we have that

$$\begin{aligned} \frac{\text{Var}(\widehat{y}|\boldsymbol{x})}{\sigma^2} &= \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} \\ &= (1, \boldsymbol{d}') \begin{bmatrix} 1/n & \boldsymbol{b}' \\ \boldsymbol{b} & \boldsymbol{C} \end{bmatrix} \begin{pmatrix} 1 \\ \boldsymbol{d} \end{pmatrix} \\ &= 1/n + 2\boldsymbol{d}'\boldsymbol{b} + \boldsymbol{d}'\boldsymbol{C}\boldsymbol{d} \end{aligned} \tag{6.24}$$

Since $\boldsymbol{C}$ is positive definite, we can simply find the point of minimum variance, $\boldsymbol{d}^*$, by taking the first partial derivatives and equate them to zero:

$$\frac{\partial \text{Var}(\widehat{y}|\boldsymbol{x})/\sigma^2}{\partial \boldsymbol{d}} = 2\boldsymbol{b} + 2\boldsymbol{C}\boldsymbol{d} = \boldsymbol{0}$$

from where

$$\boldsymbol{d}^* = -\boldsymbol{C}^{-1}\boldsymbol{b}. \tag{6.25}$$

Note how if the experimental design is orthogonal, then $\boldsymbol{d}^* = \boldsymbol{0}$ (the origin in coded units) since $\boldsymbol{b} = \boldsymbol{0}$, and in such case $\text{Var}(\widehat{y}|\boldsymbol{x})/\sigma^2 = 1/n$, in accordance with our discussion of variance properties of factorial designs.

A $100(1-\alpha)\%$ lower confidence interval on the expected response (see Appendix A) is given by

$$\widehat{y}_{min}(\boldsymbol{x}) = \widehat{y}(\boldsymbol{x}) - t_{n-p,\alpha}\, \widehat{\sigma}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}$$

where $\widehat{\sigma} = SS_{error}/(n-p)$. The "Adapted Steepest Ascent" (ASA) procedure proposed by Kleijnen et al. [81] suggests to choose the next point $\boldsymbol{d}$ in $\boldsymbol{x}' =$

$(1, \boldsymbol{d})$ such that we

$$\max_{\boldsymbol{d}} \ \widehat{y}_{\min}(\boldsymbol{x}) = \max_{\boldsymbol{d}} \ \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_0' \boldsymbol{d} - t_{n-p,\alpha}\, \widehat{\sigma} \sqrt{1/n + 2\boldsymbol{d}'\boldsymbol{b} + \boldsymbol{d}'\boldsymbol{C}\boldsymbol{d}} \quad (6.26)$$

where $\widehat{\boldsymbol{\beta}}_{-0}$ is the vector of all $\widehat{\beta}$'s except of the intercept. It turns out (see Problem 15) that $\widehat{y}_{\min}(\boldsymbol{x})$ is concave in $\boldsymbol{d}$. Therefore,

$$\frac{\partial \widehat{y}_{\min}(\boldsymbol{x})}{\partial \boldsymbol{d}} = \widehat{\boldsymbol{\beta}}_{-0} - \frac{t_{n-p,\alpha}\, \widehat{\sigma}}{\sqrt{1/n + 2\boldsymbol{d}'\boldsymbol{b} + \boldsymbol{d}'\boldsymbol{C}\boldsymbol{d}}} (\boldsymbol{b} + \boldsymbol{C}\boldsymbol{d}) = \boldsymbol{0} \quad (6.27)$$

If the experimental design is orthogonal, then $\boldsymbol{b} = \boldsymbol{0}$ and $\boldsymbol{C} = \boldsymbol{I}/n$, so the previous system of equations reduce to

$$\sqrt{\frac{1}{n}(1 + \boldsymbol{d}'\boldsymbol{d})}\, \widehat{\boldsymbol{\beta}}_{-0} - t_{n-p,\alpha}\, \frac{\widehat{\sigma}}{n} \boldsymbol{d} = \boldsymbol{0} \quad (6.28)$$

This is a system of k non-linear equations with k unknowns. It could be solved with a nonlinear solver [6]. Instead, Kleijnen et al. [81] postulate a solution with the form

$$\boldsymbol{d}^* = -\boldsymbol{C}^{-1}\boldsymbol{b} + \lambda \boldsymbol{C}^{-1} \widehat{\boldsymbol{\beta}}_{-0} \quad (6.29)$$

where the first term is equal to (6.25) and the second term adds a multiple of the estimated steepest ascent direction deflected by the variance-covariance matrix. The multiple λ is then the step size we take as we move from the point given by (6.25), which gives minimum variance (the origin, in the orthogonal case). Any point satisfying (6.27) solves problem (6.26) since $\widehat{y}_{\min}(\boldsymbol{x})$ is concave in $\boldsymbol{d}$. We then find the value of λ which makes $\boldsymbol{d}^*$ satisfy (6.27). Note how in the non-orthogonal case, $\boldsymbol{d}^*$ and λ define a direction and step size, respectively, whereas in the orthogonal case the direction is just the steepest ascent direction. We only consider the orthogonal case in what follows.

Substituting (6.29) into (6.28) we get, after some algebra

$$\left(1/n + \lambda^2 (n\, \widehat{\boldsymbol{\beta}}_{-0}' \widehat{\boldsymbol{\beta}}_{-0} - t_{n-p,\alpha}^2\, \widehat{\sigma}^2)\right) \widehat{\boldsymbol{\beta}}_{-0}' \widehat{\boldsymbol{\beta}}_{-0} = \boldsymbol{0}$$

from where

$$\lambda = \sqrt{\frac{1/n}{t_{n-p,\alpha}^2\, \widehat{\sigma}^2 - n\, \widehat{\boldsymbol{\beta}}_{-0}' \widehat{\boldsymbol{\beta}}_{-0}}} \quad (6.30)$$

[6] Such as MATLAB's `lsqnonlin`.

which gives the step size in the direction of steepest ascent. Note (see Problem 16) that (6.30) can be written as

$$\lambda = \sqrt{\frac{1/n}{t^2_{n-p,\alpha}\, MS_{error} - SS_{Regression}}} \tag{6.31}$$

where MS_{error} denotes the mean square error and SS_{Reg} denotes the sum of squares due to the regression (see Appendix A). Perez-Salvador and O'Reilly [123] point out that this quantity is well defined only when the denominator is greater than zero, which occurs only when

$$t^2_{n-p,\alpha} \; > \; \frac{SS_{Reg}}{MS_{error}}.$$

This inequality can certainly *not* hold in practice. These authors suggest to use instead a Sheffé-like critical value in the denominator of (6.31), that is, use a denominator equal to

$$p\, F_{p,n-p,\alpha} MS_{error} - SS_{Reg} \tag{6.32}$$

for some α such that

$$\frac{k+1}{k} F_{k+1,n-p,\alpha} > \frac{MS_{Reg}}{MS_{error}} \tag{6.33}$$

(recall that $p = k + 1$ for a first order model). To find an α such that this last inequality is always true, making λ real always, consider the rule for rejecting $H_o : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ at level α (see Appendix A)[7]:

$$F_{k,n-p,\alpha} \leq \frac{MS_{Reg}}{MS_{error}}. \tag{6.34}$$

Since

$$F_{k,n-p,\alpha} < \frac{k+1}{k} F_{k+1,n-p,\alpha} \quad \forall \alpha$$

then choosing α equal to the p-value of (6.34), say α^*, we get

$$F_{k,n-p,\alpha^*} \; = \; \frac{MS_{Reg}}{MS_{error}} \; < \; \frac{k+1}{k} F_{k+1,n-p,\alpha^*}$$

[7]This is the "Significance of Regression" test for a first order model, which is testing for significance of any of the derivatives of the function with respect to the factors.

and from the right hand side equality and (6.33) we see that λ obtained this way will always be positive.

This suggestion is not without problems. If the p-value α^* for the significance of regression test is large, it means no derivative is significant, and no steepest ascent search should be taken. However, as α^* increases we see that $F_{p,n-p,\alpha^*}$ decreases, and this will make the step size λ increasingly *larger*, contrary to our intuition that tells us to take cautious steps if the model fits poorly.

We now illustrate this "modified ASA" procedure with a numerical example, and show afterwards a second, different and better proposal by Perez-Salvador and O'Reilly on how to select the step size in SA.

Example. Modified Adapted Steepest Ascent. Consider the Chemical Process experiment in Section 2-2. There, recall we had that a 2^2 factorial ($n = 9$) with center runs used to fit a first order model for steepest ascent. The estimated parameters are $\boldsymbol{\beta}' = (39.57, -1.12925, 11.14)$ (the fit is not great, $R^2 = 0.6580$) and we found that $MS_{error} = 44.95$ with $n - p = 6$ dof and $MS_{regression} = 251.65$ with $p - 1 = k = 2$ dof. Using (6.30), we get

$$\lambda = \sqrt{\frac{1/9}{(4.54^2)(44.95) - 9(125.77)}}$$

and the denominator is negative. Using (6.34), the p-value is such that

$$F_{2,6,\alpha^*} = \frac{MS_{regression}}{MS_{error}} = 5.598$$

or $\alpha^* = 0.0425$. The alternative denominator (6.32) is therefore

$$3(5.1532)(44.95) - 503.30 = 191.62$$

and therefore $\lambda = \sqrt{1/9/191.62} = 0.02408$. Therefore, the next run should be conducted at the point

$$\boldsymbol{d} = \lambda \boldsymbol{C}^{-1} \widehat{\boldsymbol{\beta}}_{-0} = 9(0.02408) \begin{pmatrix} -1.29 \\ 11.14 \end{pmatrix} = \begin{pmatrix} -0.28 \\ 2.41 \end{pmatrix}.$$

This point corresponds to a somewhat larger step than the first two steps taken in Table 2.3. ■

6.4 Avoiding Extrapolation in Steepest Ascent Searches**

Perez-Salvador and O'Reilly [123] suggest[8] a different method for selecting the step size in steepest ascent by considering how much we extrapolate in the first experiment on the SA direction. Their suggested step size λ is such that

$$\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} = 0.5 \tag{6.35}$$

where

$$\boldsymbol{x}' = \left(1, \lambda \frac{\widehat{\boldsymbol{\beta}}'_{-0}}{||\widehat{\boldsymbol{\beta}}_{-0}||}\right)$$

which is based on a result by O'Reilly [121], who shows how for a point $\boldsymbol{x}$ such that $\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} \leq 1$ there exists an unbiased estimate of the CDF of the response. In such case it is said that the extrapolation is valid, otherwise the extrapolation is considered "too far away" or invalid. The value 0.5 is arbitrary, and represents a value "half way" from the center of the experimental design region to its boundaries. If the model fit is suspect, a smaller value may be used.

Example. To illustrate the computation of the step size using the idea of extrapolation, consider again the Chemical process in Section 2-2. We have that

$$\frac{\widehat{\boldsymbol{\beta}}_{-0}}{||\widehat{\boldsymbol{\beta}}_{-0}||} = \begin{pmatrix} -0.1150 \\ 0.9933 \end{pmatrix}$$

and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{9}\boldsymbol{I}_3$. Therefore, the criterion (6.35) is

$$\left(1, \lambda \frac{\widehat{\boldsymbol{\beta}}'_{-0}}{||\widehat{\boldsymbol{\beta}}_{-0}||}\right)\left(1, \lambda \frac{\widehat{\boldsymbol{\beta}}'_{-0}}{||\widehat{\boldsymbol{\beta}}_{-0}||}\right)' = \frac{9}{2} = 4.5$$

or $1 + (0.1323 + 0.98664)\lambda^2 = 4.5$ from which $\lambda = 1.76$. The next point along the SA direction is therefore

$$1.76 \begin{pmatrix} -0.1150 \\ 0.9933 \end{pmatrix} = \begin{pmatrix} -0.2024 \\ 1.7482 \end{pmatrix}$$

which corresponds to less than the first two steps taken in Table 2.3. ■

We point out that this method can be used in conjunction with any of the stopping rules for steepest ascent discussed earlier in this chapter.

[8]This section contain relatively more advanced material and may be skipped on a first reading.

6.5 Problems

1 Consider Problem 1 in Chapter 2. Find the fraction of directions excluded by the 95% confidence cone of steepest ascent.

2 Consider Problem 2 in Chapter 2. Find the fraction of directions excluded by the 95% confidence cone of steepest ascent.

3 The following response measurements have been observed along the direction of steepest *ascent*:

t	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$Y(t)$	10.0	3.89	39.39	27.17	44.15	51.41	85.33	97.23	42.96	81.23	58.32	33.42	46.25	43.36

If the Myers-Khuri stopping rule were applied to these data, what would be the stopping point and the optimal point returned? Use $\kappa = 10$, $\sigma_\varepsilon = 20$. Show, for each step, what decision is taken.

4 Repeat the previous exercise using the following response measurements observed along the direction of steepest *ascent*:

t	0	1	2	3	4	5	6	7	8	9	10	11	12
$Y(t)$	50.0	34.16	79.64	104.93	103.66	126.99	115.54	170.74	143.41	110.29	174.07	158.53	63.99

Use $\kappa = 12$, $\sigma_\varepsilon = 15$. For each step, show what decision is taken for the hypothesis being tested.

5 Apply the recursive parabolic stopping rule (RPR) to the data in Problem 3. Use $\widehat{\theta}_0 = 10, \widehat{\theta}_1 = 5, t_{prior} = 10$, and $P_0 = 10$.

6 Apply the recursive parabolic stopping rule (RPR) to the data in Problem 4. Use $\widehat{\theta}_0 = 50, \widehat{\theta}_1 = 5, t_{prior} = 12$, and $P_0 = 10$.

7 Apply the enhanced recursive parabolic stopping rule (ERPR) to the data in Problem 3. Use $\widehat{\theta}_0 = 10, \widehat{\theta}_1 = 5, t_{prior} = 10$, and $N = 7$ (window size).

8 Apply the enhanced recursive parabolic stopping rule (ERPR) to the data in Problem 4. Use $\widehat{\theta}_0 = 50, \widehat{\theta}_1 = 5, t_{prior} = 12$, and $N = 7$ (window size).

9 Consider the first order model: $\widehat{y} = 52.1 - 3.1x_1 + 6.4x_2 - 1.25x_3$, where the variance of the parameter estimates, s_b^2, equals 0.4 computed based on 5 replications of the center point of the design.

a) Does the point $\boldsymbol{x}' = (-0.9, 1.0, -0.3)$ generate a direction vector inside the 95% confidence cone of steepest ascent?

b) Find the fraction of directions excluded by the 95% confidence cone around the steepest ascent direction.

10 Suppose the following are the observed responses along different steps taken on the direction of steepest ascent:

51.4	54.0	56.8	60.9	62.1	64.2	66.1	69.0	68.2	69.9	69.8	69.9	70.9	69.2	68.3	68.0	66.0

If you were to apply the Myers-Khuri stopping rule to these data, show how would it work, i.e., which hypothesis are being tested at each step and what are the results (t_{opt} and t_{stop}) of the stopping rule. Use $\kappa = 25$ and $\sigma_\varepsilon = 1$ and assume normally-distributed errors.

11 Consider the "grade A" data in Problem 4, Chapter 2. Find the percentage of directions excluded by the 95% confidence cone in the direction of steepest ascent.

12 Consider the "grade B" data in Problem 5, Chapter 2. Find the percentage of directions excluded by the 95% confidence cone in the direction of steepest ascent.

13 Consider the "grade A" data in Problem 4, Chapter 3. Apply the modified ASA procedure described in Section 6.4 to find the step size and direction of search.

14 Consider the "grade B" data in Problem 5, Chapter 3. Apply the modified ASA procedure described in Section 6.4 to find the step size and direction of search.

15 Show that $\widehat{y}_{\min}(\boldsymbol{x})$ is concave in $\boldsymbol{d}$.

16 Show that (6.30) reduces to (6.31).

17 Write a spreadsheet function or computer program that implements the computations in the ERPR method. Your program should allow the user to apply the ERPR for any window size using Table 6.3.

Chapter 7

STATISTICAL INFERENCE IN SECOND ORDER RSM OPTIMIZATION

Everything should be made as simple as possible, but not simpler.

—Albert Einstein (1879–1955)

We continue in this chapter the discussion of methods for dealing with sampling variability in experimental optimization techniques. This chapter considers the effect of statistical sampling error in RSM techniques that are based on second order (quadratic) polynomial models. We first discuss finding confidence intervals for the eigenvalues of the Hessian matrix, that is, the effect of sampling variability in *canonical analysis*. Later sections consider the related and important problem of finding a confidence region for the optimal operating conditions $\boldsymbol{x}_0$. The unconstrained case is discussed first after which methods for the computation and display of confidence regions on constrained optima are discussed. Any traditional (frequentist) RSM optimization analysis should probably always include such regions.

7.1 Confidence Intervals on the Eigenvalues of the Matrix of Quadratic Parameter Estimates

As said earlier, any inference based on sample data is subject to sampling variability. When fitting a second order model of the form

$$\widehat{y} = b_0 + \boldsymbol{b}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$$

the eigenvalues of $\boldsymbol{B}$ determine the nature of $\widehat{y}$. This is exploited in Canonical (and Ridge) Analysis. The eigenvalues are in fact only *point estimates* of

the eigenvalues of the matrix of quadratic parameters in the "true" quadratic response surface model (here we will assume no model bias). This is relevant in practice, because an eigenvalue of $\boldsymbol{B}$ which has a large standard error is likely to *cover* the value of zero. If this is the case, we cannot claim the stationary point is *statistically* a maximum or minimum. Statistically, it would be a saddle function.

It is therefore of importance to estimate the standard errors of the eigenvalues of $\boldsymbol{B}$ and with these compute confidence intervals on each λ_i. If none of these intervals covers zero and all are to the left of zero, then there is statistical evidence that the stationary point is a maximum; if all the intervals are to the right of zero there is evidence to claim $\boldsymbol{x}_0$ is a minimum; and if some interval i covers zero this is evidence the stationary point is a saddle point.

Bisgaard and Ankenman [11] proposed a simple procedure to estimate the standard errors of the eigenvalues. Their method suggests to fit first the quadratic polynomial and compute the eigenvectors of $\boldsymbol{B}$ from $\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M} = \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix containing the eigenvalues and $\boldsymbol{M}$ is an orthonormal matrix of eigenvectors of $\boldsymbol{B}$ (see Appendix C). Next, we get the rotated coordinates without translation. That is, define $\boldsymbol{u} = \boldsymbol{M}'\boldsymbol{x}$ or $\boldsymbol{x} = \boldsymbol{M}\boldsymbol{u}$ for a given point $\boldsymbol{x}$. With this transformation, the resulting model can be easily shown to be (see Problem 1):

$$\widehat{y} = \widehat{y}(\boldsymbol{u}) = a_0 + \boldsymbol{a}'\boldsymbol{u} + \sum_{i=1}^{k} \widehat{\lambda}_i u_i^2$$

which is sometimes called the "A-canonical form" [20] of the quadratic model (it is easy to see that $a_0 = b_0$ and $\boldsymbol{a} = \boldsymbol{M}'\boldsymbol{b}$). This is just another linear regression model in the u's, which can be fit using ordinary least squares. For an experimental design matrix $\boldsymbol{A}$ ($n \times k$ matrix corresponding to the k factors) after transformation we get the rotated design matrix $\boldsymbol{U} = [\mathbf{1}|\boldsymbol{A}\boldsymbol{M}|quad]$ where "quad" are the k columns corresponding to the pure quadratic terms obtained from the k columns that contain the rotated main effects u_i's, and $\mathbf{1}$ is a column of ones. The standard errors of the eigenvalues, $std(\widehat{\lambda}_i)$, can therefore be obtained from the corresponding $\widehat{\sigma}^2(\boldsymbol{U}'\boldsymbol{U})^{-1}$ matrix, with $\widehat{\sigma}^2$ obtained from the sum of squared errors of this second fit, which only has $p = 2k + 1$.

The $100(1-\alpha)\%$ confidence interval on eigenvalue i is given by

$$\widehat{\lambda}_i \pm t_{\alpha/2,n-p} std(\widehat{\lambda}_i), \quad i = 1, \ldots, k$$

The confidence may be adjusted via Bonferroni's inequality to account for the effect of the k simultaneous confidence intervals computed.

Example. A 3^3 experimental design ($n = 27$) is run to fit a full quadratic polynomial. The resulting fitted model is

$$\widehat{y} = 23.17 + 3.232x_1 + 4.005x_2 - 3.311x_3 + 0.424x_1x_3 + 2.994x_1x_3 - 4.674x_2x_3 + 1.582x_1^2 + 3.041x_2^2 + 4.537x_3^2$$

with $\widehat{\sigma}^2 = 6.1939$. The matrix of estimated quadratic coefficients is therefore:

$$\boldsymbol{B} = \begin{bmatrix} 1.582 & 0.2121 & 1.4972 \\ 0.2121 & 3.041 & -2.3373 \\ 1.4972 & -2.3373 & 4.537 \end{bmatrix}$$

and thus, the eigenvalues are $\widehat{\lambda_1} = 0.30432$, $\widehat{\lambda_2} = 2.3607$, and $\widehat{\lambda_3} = 6.4934$. This seems to indicate that the stationary point $\boldsymbol{x} = 0.5\boldsymbol{B}^{-1}b = [2.7649, -0.8542, -1.7174]'$ (in coded units) is a point of minimum response.

However, the sampling variability of the eigenvalues needs to be taken into account. The matrix $\boldsymbol{M}$ of eigenvectors[1] that diagonalizes matrix $\boldsymbol{B}$ is

$$\boldsymbol{M} = \begin{bmatrix} -0.69358 & 0.68447 & 0.22463 \\ 0.49815 & 0.68094 & -0.53681 \\ 0.52039 & 0.26042 & 0.81325 \end{bmatrix}$$

The $\boldsymbol{A}$ matrix corresponds to the three columns of a 3^3 designs in coded factors (a 27×3 matrix). Thus, $\boldsymbol{U} = [\mathbf{1}|\boldsymbol{AM}|quad]$ is a 27×7 matrix, and

$$\widehat{\sigma}^2(\boldsymbol{U}'\boldsymbol{U})^{-1} = \left[\begin{array}{ccccccc} 1.3649 & 0.0000 & 0.0000 & 0.0000 & -0.5850 & -0.5850 & -0.5850 \\ \hline 0.0000 & 0.2925 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2925 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.2925 & 0.0000 & 0.0000 & 0.0000 \\ \hline -0.5850 & 0.0000 & 0.0000 & 0.0000 & \mathbf{0.4450} & 0.2254 & 0.2072 \\ -0.5850 & 0.0000 & 0.0000 & 0.0000 & 0.2254 & \mathbf{0.4608} & 0.1914 \\ -0.5850 & 0.0000 & 0.0000 & 0.0000 & 0.2072 & 0.1914 & \mathbf{0.4790} \end{array}\right]$$

Therefore, the standard errors of the three estimated eigenvalues are 0.4450, 0.4608, and 0.4790. The 95% simultaneous confidence intervals for the eigenvalues (neglecting Bonferroni effects) are[2]:

$$\widehat{\lambda_1} = 0.3043 \pm 2.4213(0.4450) = 0.3043 \pm 1.0782$$

[1]This is easy to compute with Matlab's `eig` function.
[2]Note that $t_{0.025,20} = 2.4213$.

and, similarly,

$$\widehat{\lambda_2} = 2.3607 \pm 1.1163$$

and

$$\widehat{\lambda_3} = 6.4939 \pm 1.1606$$

and therefore, since the first interval covers zero, there is evidence that the fitted response is a saddle function in reality, and that the stationary point corresponds to a saddle point. ■

The method of Bisgaard and Ankenman assumes all eigenvalues are distinct. The possibility of repeated eigenvalues, although remote in RSM practice, has been treated by Yin and Seymour [159].

7.2 Confidence Region on the Location of a Stationary Point

Suppose[3] we fit

$$\widehat{y} = b_0 + \boldsymbol{b}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}.$$

To get the stationary point we compute

$$\frac{\partial \widehat{y}}{\partial \boldsymbol{x}} = \boldsymbol{d}(\boldsymbol{x}) = \boldsymbol{b} + 2\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$$

where $\boldsymbol{d}(\boldsymbol{x})$ is the first derivative vector. If we run the DOE again, and fit the model from new experimental data, the location of the stationary point obtained this way will vary. In other words, the stationary point $\boldsymbol{x}_0 = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b}$ is simply a *point estimate* of the true stationary point of a hypothetical "true" quadratic response function, assuming no model bias. In analogy to basic statistical procedures, a confidence region on the location of the stationary point will help us determine a region in the controllable factor space, which, if computed several times after many experiments would cover the true stationary point some proportion of times.

Confidence regions on stationary points are in principle useful in two ways:

1 to determine how precise our point estimate $\boldsymbol{x}_0$ is;

2 to propose a region where alternative operating conditions exist that provide a solution that does not differ significantly from the response at the

[3]This section is based on reference [41].

estimated stationary point, in case the stationary point solution is impractical for reasons not modeled in the DOE (e.g., cost of running the process at $\boldsymbol{x}_0$).

A procedure for obtaining such regions was presented by Box and Hunter [23]. To obtain the confidence region we proceed as follows. It is easy to see that the sample variability in the parameter estimates will be transmitted to the derivative vector, and this in turn will be transmitted to the stationary point $\boldsymbol{x}_0$. At the stationary point we have that

$$\boldsymbol{d}(\boldsymbol{x}) \sim N\left(\boldsymbol{0}, \text{Var}(\boldsymbol{d}(\boldsymbol{x}))\right)$$

thus, the vector of derivatives follows a multivariate normal distribution with mean zero and some variance covariance matrix of the form

$$\text{Var}(\boldsymbol{d}(\boldsymbol{x})) = \sigma^2 \boldsymbol{V}$$

where $\boldsymbol{V}$ is a $k \times k$ matrix that contains the variances and covariances of the derivatives of $\widehat{y}$. The quantity $\boldsymbol{d}(\boldsymbol{x})' \text{ Var}(\boldsymbol{d}(\boldsymbol{x}))^{-1} \boldsymbol{d}(\boldsymbol{x})$ is a quadratic form of multivariate normal variables with mean zero. Therefore, it is distributed as

$$\boldsymbol{d}' \text{ Var}(\boldsymbol{d})^{-1} \boldsymbol{d} = \frac{\boldsymbol{d}' \boldsymbol{V}^{-1} \boldsymbol{d}}{\sigma^2} \sim \chi^2_k$$

(omitting the dependency on $\boldsymbol{x}$), a (central) chi-square distribution with k degrees of freedom. We also know that

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p},$$

where $\boldsymbol{d}$ and s^2 are independent. From the definition of an F random variable,

$$\frac{\frac{\boldsymbol{d}' \boldsymbol{V}^{-1} \boldsymbol{d}}{\sigma^2} / k}{\frac{(n-p)s^2}{\sigma^2} / (n-p)} = \frac{\boldsymbol{d}' \boldsymbol{V}^{-1} \boldsymbol{d}}{k s^2} = \frac{\boldsymbol{d}' \, \widehat{\text{Var}}(\boldsymbol{d}(\boldsymbol{x}))^{-1} \boldsymbol{d}}{k} \sim F_{k,n-p}. \tag{7.1}$$

Therefore,

$$P\left\{ \boldsymbol{d}'(\boldsymbol{x}) \widehat{\text{Var}}(\boldsymbol{d}(\boldsymbol{x}))^{-1} \boldsymbol{d}(\boldsymbol{x}) \leq k F_{\alpha,k,n-p} \right\} = 1 - \alpha$$

where $\widehat{\text{Var}}(\boldsymbol{d}(\boldsymbol{x})) = s^2 \boldsymbol{V}$. Vectors $\boldsymbol{x}$ satisfying the quantity inside the curly braces are inside the $100(1-\alpha)\%$ confidence region for the stationary point, $\boldsymbol{x}$. The stationary point estimate $\widehat{\boldsymbol{x}_0} = -\frac{1}{2} \boldsymbol{B}^{-1} \boldsymbol{b}$ will be inside this region. To *plot*

the region is cumbersome (up to recently, only 2-dimensional BH confidence regions had been plotted). One has to draw all points $\boldsymbol{x}$ such that the inequality inside the curly braces is an equality. This will give the edge of the region.

A MAPLE computer program that plots the Box-Hunter confidence regions up to $k = 5$ was written by S. Cahya and is described by del Castillo and Cahya [41]. One of the findings in this paper is that the BH confidence regions may be composed of disjoint subsets. The reason for this is that the resulting confidence regions are for *all stationary points that could occur through repeated sampling*, regardless of whether such stationary points are a minimum, maximum or saddle point. For instance, it can happen that the function is not significantly convex statistically speaking, so in repeated sampling, its parameter estimates will result in a mix of minimums and saddle points. This will tend to happen when at least one of the eigenvalues of the $\boldsymbol{B}$ matrix has a confidence interval that covers zero at some confidence level. If this is the case, the saddle points may be located far away from the minimums, forming two disjoint regions. An example from Box and Draper [20] where exactly this happens is shown below.

One has to keep in mind that the term $\boldsymbol{d}(\boldsymbol{x})' \widehat{\text{Var}}(\boldsymbol{d}(\boldsymbol{x}))\boldsymbol{d}(\boldsymbol{x})$ is not a paraboloid but a fourth order polynomial in $\boldsymbol{x}$. This will result in very complicated shapes of the confidence region, which will be hard to interpret. The size of the confidence region depends on two aspects:

1 how flat the true function is in the vicinity of the stationary point (the flatter, the larger the region);

2 how well the response model fits (the worse fit the larger the region).

Thus, it is not true that for a stationary point that is maximum or a minimum, a large confidence region is bad news. A flat region gives flexibility to process engineers when choosing operating conditions for the process. Thus, model checks should complement any conclusions from such a confidence region in order to discern the reason for a confidence region that is large relative to the experimental region.

Example. Computation of a Box-Hunter confidence region. Consider the three factor experiment reported by Box and Draper [20, p. 305]. In this experiment, the goal was to find the percentage concentration of two constituents (x_1 and x_2, in coded units) and the temperature (x_3, coded) that

Table 7.1. Experimental design data (from [20])

x_1	x_2	x_3	y
-1	-1	-1	25.74
1	-1	-1	48.98
-1	1	-1	42.78
1	1	-1	35.94
-1	-1	1	41.50
1	-1	1	50.10
-1	1	1	46.06
1	1	1	27.70
-2	0	0	35.50
2	0	0	44.18
0	-2	0	38.58
0	2	0	28.46
0	0	-2	33.50
0	0	2	42.02
0	0	0	57.52
0	0	0	59.68

maximize the elasticity of certain polymer. The experimental design and data are shown in Table 7.1.

The second order polynomial model fit is

$$\begin{aligned} \widehat{y} \;=\; & 57.31 + 1.5x_1 - 2.13x_2 + 1.81x_3 \\ & -4.69x_1^2 - 6.27x_2^2 - 5.21x_3^2 \\ & -7.13x_1x_2 - 3.27x_1x_3 - 2.73x_2x_3 \end{aligned} \tag{7.2}$$

with stationary point $\widehat{\boldsymbol{x}^*} = (0.4603, -0.4644, 0.1509)'$, which, from the eigenvalues of matrix $\boldsymbol{B}$ is a maximum point.

The fitted full quadratic model has an excellent fit, with $R^2 = 0.972$ (adj $R^2 = 0.930$), no lack of fit, all terms in the model are statistically significant, and the usual regression diagnostics do not show any abnormality. Box and Draper [20, p. 339] discuss how the stationary point corresponds to a maximum point given the eigenvalues and approximate standard errors of the eigenvalues of $\boldsymbol{B}$. They describe how there is elongation in the response in one of the

canonical directions, clearly seen from a contour plot of $\widehat{Y}$ vs. x_1 and x_2 (see Figure 7.1).

Box and Draper indicate that in this example

> A more accurate picture about what is known at this stage of experimentation can be gained by the confidence region calculation described in Box and Hunter, 1954.

but provided no such calculation. Let us suppose that a 90% confidence region is desired.

The $\boldsymbol{d}(\boldsymbol{x})$ vector is

$$\begin{bmatrix} 1.5000 - 9.3800\, x_1 - 7.1300\, x_2 - 3.2700\, x_3 \\ -2.1300 - 7.1300\, x_1 - 12.5400\, x_2 - 2.7300\, x_3 \\ 1.8100 - 3.2700\, x_1 - 2.7300\, x_2 - 10.4200\, x_3 \end{bmatrix}$$

with a scaled variance-covariance matrix $\boldsymbol{V}$ equal as shown in Table 7.2.

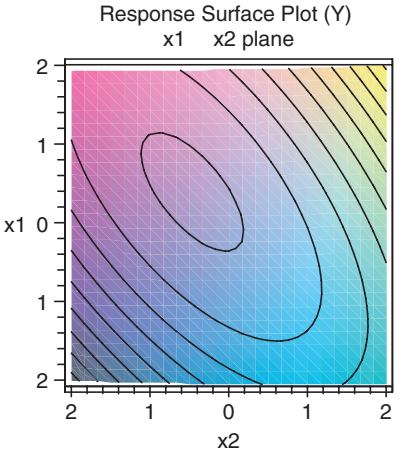


Figure 7.1. Contour plot of fitted response vs. x_1 and x_2. Source: [41]

Table 7.2. The scaled variance-covariance matrix $\boldsymbol{V}$ in the example (from [41])

$$\begin{bmatrix} 0.432 + 1.731\, {x_1}^2 + 0.865\, {x_2}^2 + 0.865\, {x_3}^2 & 1.731\, x_1\, x_2 & 1.731\, x_1\, x_3 \\ 1.731\, x_1\, x_2 & 0.432 + 0.865\, {x_1}^2 + 1.731\, {x_2}^2 + 0.865\, {x_3}^2 & 1.731\, x_2\, x_3 \\ 1.731\, x_1\, x_3 & 1.731\, x_2\, x_3 & 0.432 + 0.865\, {x_1}^2 + 0.865\, {x_2}^2 + 1.731\, {x_3}^2 \end{bmatrix}.$$

It is immediately obvious that even in this relatively simple case, manually obtaining the expression for $\boldsymbol{d}'\boldsymbol{V}^{-1}\boldsymbol{d}$ is extremely tedious. The aforementioned MAPLE program can be used for this purpose, and then used to graph and display the confidence region.

Figure 7.2 shows the 90% confidence region plots for this experiment. Even though the fit of the quadratic model is excellent, the confidence region is not only open, but *disconnected*. Such a region occurs because some parameter values within the confidence region for the parameters in the $\boldsymbol{\Xi}$ matrix result in saddle functions, and generate the stationary points in the smaller sub-region, far from the stationary points in the larger sub-region which contains only maxima of concave functions. The important message is that the Box-Hunter

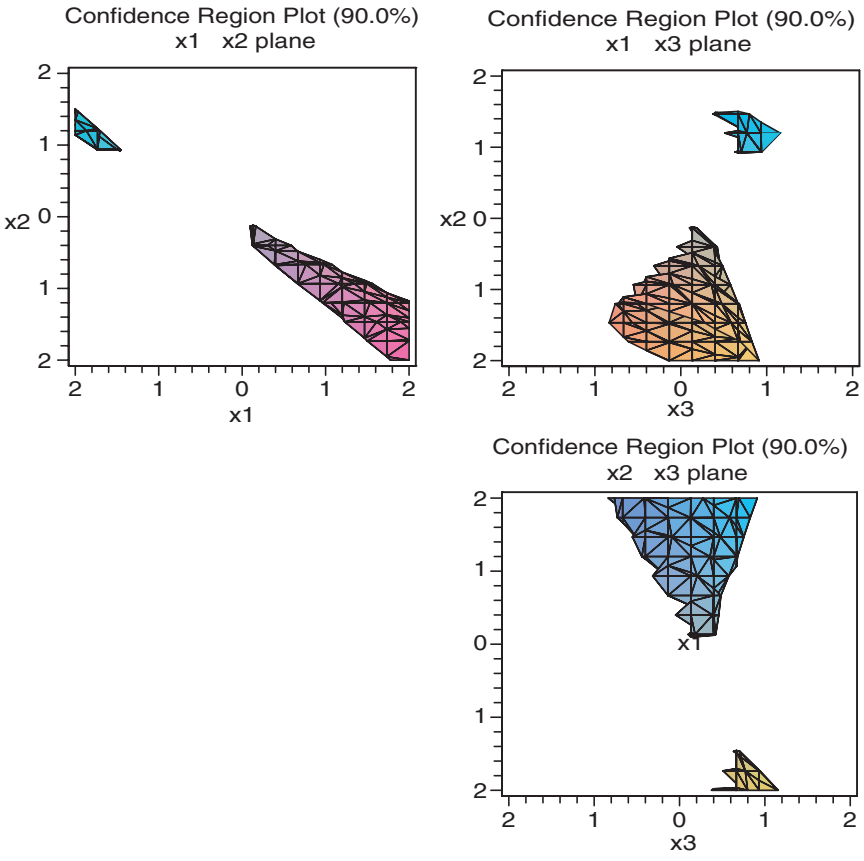


Figure 7.2. 90% Box-Hunter confidence region plots for the stationary points in the example. The stationary point of the fitted model is (0.4603, −0.4644, −0.1509). From: [41]

technique gives a confidence region for *all types* of stationary points which in turn originate from the confidence region of the parameter values. This includes true maxima and minima *and* saddle points. A more exact analysis of the eigenvalues in this problem, using the procedure suggested in Peterson [124] with a critical value of $c_\alpha^2 = kF_{\alpha,k,n-q}$ (where in this case $q = 6$ is the number of parameters in $\boldsymbol{B}$) reveals that the 90% upper confidence bound for the largest eigenvalue is -0.34. However, a 95% upper bound equals 0.03. Thus there is some probability that the response is not a statistically concave quadratic function. There is a non-negligible probability that the response is a saddle. In such instances, the stationary points will be located in the smaller sub-region. ■

This example indicates a potential weakness in the Box-Hunter confidence region computation. In cases in which the fitted response is a saddle, the experimenter should use instead a technique for finding the confidence region for the true maximum of the surface subject to some constraint, as we discuss in the next section.

7.3 Confidence Regions for the Constrained Maximum or Minimum of a Response Function**

The previous discussion[4] shows that the Box and Hunter confidence regions are not very useful in practice, as they will provide a region for all stationary points, and therefore the regions may be disconnected or relate to an uninteresting saddle function. Even when the stationary point is indeed an optimal point, this point is sometimes located outside the region of operability hence the optimal solution cannot be implemented. Therefore, in practice, solutions should be limited to be inside of a constrained region of interest. Furthermore, some problems are naturally constrained, such as mixture experiments where the fraction of the mixture ingredients should add up to one.

The underlying optimization problem in Box-Hunter (BH) confidence regions is an unconstrained one, essentially the same problem addressed in "Canonical analysis". A procedure that eliminates the problems in the unconstrained BH confidence regions was proposed by Peterson, Cahya and del Castillo [126] (hereafter their procedure is referred to as the PCD method). The procedure finds a confidence region for the location of the *optimum* of

[4]This section has somewhat more advanced material and may be skipped on a first reading.

any response function model linear in the parameters subject to an arbitrary constraint that depends on $\boldsymbol{x}$ but not on any unknown parameter.

Stablein, Carter, and Wampler [146] proposed a methodology for computing the confidence regions for constrained optimal points. These authors modified the BH approach by using the Lagrange multiplier approach to incorporate the constraints. However, their method assumes the lagrangian multipliers as constants, hence ignoring the sampling variability of the multipliers. This assumption is not valid since the multipliers are dependent on the model parameter estimates and these estimates are subject to sampling variability.

Peterson et al. [126] proposed a generalized approach that has some advantages over the Stablein et al. method. This approach does not use lagrangian multipliers, so it avoids the technical difficulty of incorporating the sampling variability of the multipliers. Moreover, the method generalizes the quadratic polynomial model required in Stablein et al. (1983) into any models that are linear in the parameters. This is especially beneficial in mixture experiments where "exotic" functions of the factors sometimes provide a better fit than quadratic polynomials.

The PCD method for confidence regions of constrained optima

Let us assume that the response surface under consideration is linear in the parameters, i.e., it can be modeled as

$$y = \beta_0 + \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta} + \epsilon, \tag{7.3}$$

where y is the response variable, β_0 is the intercept term, $\boldsymbol{z}(\boldsymbol{x})$ is a $(p-1) \times 1$ vector valued function of the $k \times 1$ vector $\boldsymbol{x}$ (i.e., $\boldsymbol{z}(\boldsymbol{x})$ corresponds to $\boldsymbol{x}$ in model form), $\boldsymbol{\theta}$ is a $(p-1) \times 1$ vector of regression coefficients, and ϵ is the error term that follows a normal distribution with mean 0 and variance σ^2.

Without loss of generality, suppose that we want to minimize the response y. Let $\boldsymbol{x}_0$ and $\eta(\boldsymbol{\theta})$ be defined as

$$\boldsymbol{z}(\boldsymbol{x_0})'\boldsymbol{\theta} = \min_{\boldsymbol{x} \in \mathcal{R}} \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta} = \eta(\boldsymbol{\theta}), \tag{7.4}$$

where $\mathcal{R}$ is the constraining region (for unconstrained optimization $\mathcal{R}$ is simply the space of $\boldsymbol{x}$). Note that $\boldsymbol{x_0}$ is the true minimizer of the response surface and it is unknown. To compute the $100(1-\alpha)\%$ confidence region of $\boldsymbol{x_0}$, one can find the set of all $\boldsymbol{x}$-values such that the null hypothesis $H_0 : \eta(\boldsymbol{\theta}) - \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta} = 0$

is not rejected at level α. This hypothesis needs to be tested for all points $\boldsymbol{x}$ on a grid defined over $\mathcal{R}$. In other words, when H_0 is not rejected for a point $\boldsymbol{x}$, the point $\boldsymbol{x}$ must belong to the confidence region for the optimum since there is statistical evidence to conclude that the response value at $\boldsymbol{x}$ does not differ from the minimum response value. PCD proposed a way to compute the $100(1-\alpha)\%$ confidence interval for $\Big(\eta(\boldsymbol{\theta}) - \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta}\Big)$ and rejecting H_0 simply by checking whether this confidence interval excludes 0. Since $\Big(\eta(\boldsymbol{\theta}) - \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta}\Big) \leq 0$ for all $\boldsymbol{\theta}$, the lower bound of the confidence interval is always less than zero. Therefore, the confidence interval for $\Big(\eta(\boldsymbol{\theta}) - \boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta}\Big)$ excludes zero if the upper bound is less than zero. This upper bound is given by (see PCD for the derivation):

$$U_B = \min_{\boldsymbol{w}\in\mathcal{R}} b_x(\boldsymbol{w}), \tag{7.5}$$

where

$$b_x(\boldsymbol{w}) = \Big(\boldsymbol{z}(\boldsymbol{w}) - \boldsymbol{z}(\boldsymbol{x})\Big)'\widehat{\boldsymbol{\theta}} + \sqrt{c_\alpha^2\left[\Big(\boldsymbol{z}(\boldsymbol{w}) - \boldsymbol{z}(\boldsymbol{x})\Big)'\widehat{\boldsymbol{V}}\Big(\boldsymbol{z}(\boldsymbol{w}) - \boldsymbol{z}(\boldsymbol{x})\Big)\right]}. \tag{7.6}$$

Here, $\widehat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{V}}$ is the estimated variance-covariance matrix of $\boldsymbol{z}(\boldsymbol{x})'\boldsymbol{\theta}$, c_α^2 is the $100(1-\alpha)\%$ upper percentile of an F-distribution with k and $n-p$ degrees of freedom.

Equation (7.5) allows us to construct the $100(1-\alpha)\%$ confidence region for the optimal factor level $\boldsymbol{x_0}$ in a straightforward manner. This confidence region is denoted as $\mathcal{C}_R$ for brevity. A point $\boldsymbol{x} \in \mathcal{R}$ does not belong to $\mathcal{C}_R$ if $U_B < 0$. However, it is not necessary to perform the minimization in (7.5) and get the minimum value U_B. To discard a point $\boldsymbol{x}$ from being inside the $\mathcal{C}_R$, it is only necessary to find a single other point $\boldsymbol{w} \in \mathcal{R}$ such that the function $b_x(\boldsymbol{w}) < 0$, since $U_B \leq b_x(\boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{R}$. This clearly would imply that $U_B < 0$. Furthermore, for the specific case when the response surface satisfies the following two conditions:

(i) the response is quadratic (hence it is differentiable);

(ii) the response is strictly convex in a statistical sense, i.e. the matrix of quadratic coefficients $\boldsymbol{B}$ is positive definite (p.d.) for all $\theta \in \mathcal{C}$ where $\mathcal{C} = \{\boldsymbol{\theta} : (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\widehat{\boldsymbol{V_\theta}}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq c_\alpha^2\}$ is the $100(1-\alpha)\%$ confidence region for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{V_\theta}}$ is an estimate of $\boldsymbol{V_\theta}$, the variance covariance matrix of $\widehat{\boldsymbol{\theta}}$,

PCD proposed the following two-step derivative approach:

Step 1 (Box-Hunter (BH) Step): For all points $\boldsymbol{x}$ in the interior of $\mathcal{R}$, select confidence region points according to the BH criterion.

Step 2: For all points $\boldsymbol{x}$ on the boundary of $\mathcal{R}$, search for the first $\boldsymbol{d}$-value on $\mathcal{D}$ such that $b'_x(\boldsymbol{x};\boldsymbol{d}) < 0$, where $\mathcal{D} = \{\boldsymbol{d} : \boldsymbol{d}'\boldsymbol{d} = 1, \boldsymbol{x} + h\boldsymbol{d} \in \mathcal{R}$ for small $h > 0\}$ and $b'_x(\boldsymbol{w};\boldsymbol{d}) = \lim_{h\to 0^+}[b_x(\boldsymbol{w} + h\boldsymbol{d}) - b_x(\boldsymbol{w})]/h$.

PCD showed that, for differentiable response models, $b'_x(\boldsymbol{x};\boldsymbol{d}) = \boldsymbol{d}'\boldsymbol{D}(\boldsymbol{x})\widehat{\theta} + c_\alpha[\boldsymbol{d}'\boldsymbol{D}(\boldsymbol{x})\widehat{\boldsymbol{V}}\boldsymbol{D}(\boldsymbol{x})'\boldsymbol{d}]^{1/2}$ where $\boldsymbol{D}(\boldsymbol{x})$ is the $k \times (p-1)$ matrix of partial derivatives of $\boldsymbol{z}(\boldsymbol{x})$ with respect to $\boldsymbol{x}$. For subsequent discussions, the first step of this derivative approach will be referred to as the BH Step.

In general, some response models do not satisfy differentiability nor convexity assumptions. We note that even though a fitted quadratic polynomial is strictly convex (i.e. $\widehat{\boldsymbol{B}}$ is p.d.), $\boldsymbol{B}$ might not be p.d. everywhere on $\mathcal{C}$. For general response surface models, a two-step derivative-free approach was then proposed:

Step 1: Compute the estimated optimal point $\widehat{\boldsymbol{x}_0}$. Set $\boldsymbol{w} = \widehat{\boldsymbol{x}_0}$ and use $b_x(\widehat{\boldsymbol{x}_0})$ as a criterion to reject as many points in $\mathcal{R}$ as possible.

Step 2: For any $\boldsymbol{x}$ not rejected in step 1, search for the first $\boldsymbol{w} \in \mathcal{R}$ such that $b_x(\boldsymbol{w}) < 0$.

For response surfaces that are differentiable on $\mathcal{R}$ but not necessarily convex, an efficient algorithm can be constructed by adding the BH Step (for checking points in the interior of R) in between steps 1 and 2 of the derivative-free approach shown above.

Cahya et al. [31] provide an easy to use MATLAB program that implements an improved version of the PCD algorithm, providing a faster algorithm than the original PCD method, hereafter called the CDP method[5]. The accuracy of the CDP algorithm has the important property of being independent of the grid used in the computations (contrary to the PCD method) and applicable to differentiable and non-differentiable response models.

Example. A non-differentiable mixture problem. Consider a mixture experiment that studies the effect of three chemical substances on the glass transi-

[5]As with other programs mentioned in this book, the Matlab code, called CONREG, is available at the author's (EDC) personal web page (Engineering Statistics Laboratory, Penn State University).

tion temperature [57] was used. The objective of the study was to find the optimal factor levels associated with minimizing glass transition temperature. The chemical substances were *Pluronic F68* (x_1), *polyoxyethelene 40 monostearate* (x_2), and *polyoxyethelene sorbitan fatty acid ester NF* (x_3). The experimental design used was a modified McLean-Anderson design [94] with two centroid points, resulting in a sample size of eleven. The response surface model that gave the best fit was an H1 Becker model [6]:

$$\begin{aligned} y &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} \min(x_1, x_2) \\ &\quad + \beta_{13} \min(x_1, x_3) + \beta_{23} \min(x_2, x_3) + \epsilon, \end{aligned} \tag{7.7}$$

where y equals the observed glass transition temperature (^oC). The mean squared error (MSE) associated with this model is 1.71 which is a 53% reduction in MSE from the standard quadratic model. The $adjusted$-R^2 for the model is 96.4%. Note that the response model in (7.7) is not everywhere differentiable and is subject to the mixture constraints. Thus use of the BH (derivative based) step would not be convenient to implement, requiring careful checking of each $\boldsymbol{x}$-point to make sure that $x_i \neq x_j$ or that $\boldsymbol{x}$ is not on the boundary of mixture experimental region. Furthermore, for any mixture model, use of the BH Step requires that the $x_1 + \ldots + x_k = 1$ constraint be eliminated by reducing the model to $(k-1)$ independent factors.

The confidence regions for the optimal factor levels were computed for three different grid resolutions: 3721, 10201, and 40401 grid points, respectively, inside a unit square (the constrained region for the experiment is a simplex region which is a triangle inside a unit square). The confidence regions are plotted in Figure 7.3. The figure illustrates the original PCD region at different grid resolution and the enhanced CDP method proposed by Cahya et al. [31]. As it can be seen, the accuracy shape) of the region is independent for the latter, but not for the former. ■

While the CDP method does not *guarantee* a given coverage, extensive simulations (see [30]) have shown that is actual coverage is very close to nominal, and therefore, it is of considerable value in practice. This is particularly true given the availability of an easy-to-use computer program that implements the method.

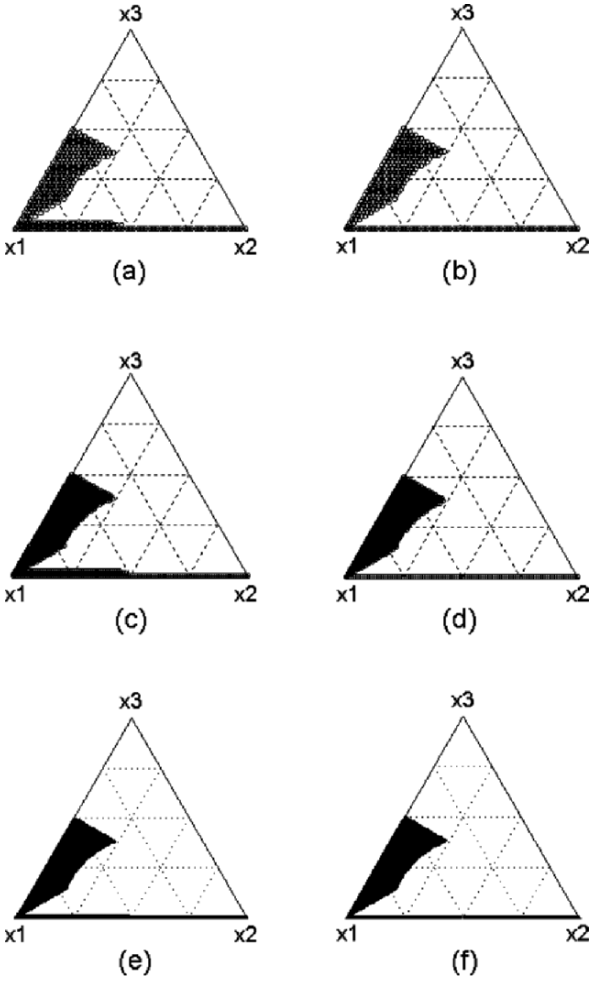


Figure 7.3. 95% Confidence Regions for Frisbee and McGinity [57] example. The confidence regions were obtained using (a) PCD approach (3271); (b) CDP approach (3271); (c) PCD approach (10201); (d) CDP approach (10201); (e) PCD approach (40401); (f) CDP (40401). The numbers in the parentheses are the total number of grid points inside a unit square. The resulting confidence regions of the PCD derivative-free approach (a,c,e) are inaccurate (bigger than they actually are). This inaccuracy is decreased as the grid resolution is increased. Source: [31]

7.4 Problems

1 Show that the resulting model after translation using $\boldsymbol{x} = \boldsymbol{M}\boldsymbol{u}$ where $\boldsymbol{M}'\boldsymbol{B}\boldsymbol{M} = \boldsymbol{D}$ with $\boldsymbol{D}$ a diagonal matrix with the eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{M}$ the matrix of corresponding eigenvectors is the "A-canonical" model.

2 Consider a full quadratic model in k factors that is to be minimized subject to $\boldsymbol{x}'\boldsymbol{x} \leq \rho^2$, just as in ridge analysis.

a) Using a Lagrange multiplier for the constraint, find an expression for the $100(1-\alpha\%)$ confidence region for the constrained optimum assuming the Lagrange multiplier μ is a known *constant* (a value of μ can certainly be obtained numerically from the problem's data, so this is not a problem).

b) What is wrong with the assumption of a constant Lagrange multiplier?

3 Use MATLAB and the CONREG program available the author personal web page (Engineering Statistics Laboratory) to build a confidence region for the stationary point of the following data.

x_1	x_2	Yield
-1	-1	88.55
1	-1	85.8
-1	1	86.29
1	1	80.4
-1.414	0	85.5
1.414	0	85.39
0	-1.414	86.22
0	1.414	85.7
0	0	91.21
0	0	91.85
0	0	91.31
0	0	91.94

Provide a) the input data file, and b) a plot of the resulting 95% confidence region in a region bounded by $(-1, -1)$ and (1,1). Use $c_\alpha^2 = F_{\alpha,k,n-p}$ and specify a "maximum".

4 Repeat the previous problem for the metal cutting experiment in Chapter 1.

5 Repeat the previous problem for the "grade A" response in Table 2.6.

6 Repeat the previous problem for the "grade B" response in Table 2.6.

Chapter 8

BIAS VS. VARIANCE

It may be remarked that nobody has yet shown that an erroneous mathematical logic is incapable of predicting phenomena verifiable by experiment.

—Eric Temple Bell (1883–1960)

From the late 1950's to up to the later 1970's a debate ensued in the Statistics community between two schools of thought in experimental design. The main point of contention was the practical utility of optimal experimental design theory, mainly developed by J. Kiefer and co-workers. We will refer to this school as the "optimal design school" in this chapter. As mentioned in Section 5.7, using optimality theory one designs an experiment that is optimal in some precisely defined way for a given model form; the design will not be optimal, and probably, not even "robust" if the true process obeys a model different than the assumed one[1]. This point of view is held by G. Box and his co-workers, which we will refer here as the "Applied Statistics school". The purpose of the present chapter is to introduce the main ideas behind this debate. Since very few practical conclusions resulted from the debate itself, the chapter is necessarily short (the methods developed by both schools have had a great impact, but that is the matter for the other chapters in this book).

For a fuller description of this debate from the point of view of the Applied Statistics school there are at least 4 books, see [112, 20, 76, 117]. Although

[1]There has been recent work where methods for optimal designs that are robust to the model form have been proposed. See, e.g., [68] and the comments in the conclusion to this chapter.

there are several books on optimal design theory, these are highly abstract and for the most part do not pay attention to the bias vs. variance debate. Perhaps the best references where Kiefer and co-workers argued in favor of their side in the Bias vs. Variance debate are the papers [78, 58]. The initial round of this "clash of statistical cultures", as H. Wynn [158] has called it[2] was during the presentation of a famous paper by Kiefer at the Royal Society [77]. This was received with scorn by some of the society members, in particular by Yates. As mentioned by Wynn, this subject pertains to the "sociology" of experimental design in the XX century. The debate pertains also to statistical inference aspects of experimental design, and it is because of this that is presented in the present part of this book.

8.1 The Applied Statistics Argument

Box and Draper [21] argued that bias is a much more important consideration than variance when designing an experiment. The argument goes as follows. Suppose that the true unknown system we want to model is given by

$$g(x_1, x_2, \ldots, x_k) = g(\mathbf{x})$$

where g is a polynomial of order d_2, but we use instead a model

$$\widehat{y}(x_1, x_2, \ldots, x_k) = \widehat{y(\mathbf{x})},$$

a polynomial of order $d_1 < d_2$. Both polynomials are defined over a region of interest R. The difference $\widehat{y} - g$ represents the discrepancy between the fitted model and the true system. These differences occur due to two causes:

1 sampling (experimental) error. This is the "variance" component of the difference or error.

2 systematic differences between $\widehat{y}$ and g. This is the "bias" component of difference or error.

To study the relative importance of these two components of the error, Box and Draper [21] proposed to use the following criterion which accounts for both types of error above:

$$J = \frac{n}{\sigma^2} \frac{\int_R E[\widehat{y(\mathbf{x})} - g(\mathbf{x})]^2 d\mathbf{x}}{\int_R d\mathbf{x}}. \tag{8.1}$$

[2] In a comment to Kiefer's paper.

This is the expected squared deviation over the region[3] R, normalized with respect to the volume of R, the variance σ^2, and the number of observations in the experimental design n. An assumption in this criterion, to which we will return when we show the opposing view in this debate[4], is that the expected value assumes a uniform distribution of the x's over R, that is, all points inside R are considered equally important as they receive the same weight in (8.1). In turn, non-uniform distributions were further considered by Box and Draper [20].

Box and Draper suggested to evaluate J for designs whose coordinates are standardized according to

$$x_{iu} = \frac{\xi_{iu} - \overline{\xi}_i}{s_i}$$

where ξ_{iu} denotes the factor in original units and s_i denotes the standard deviation of factor i (using a denominator of n) in the design (see Section 5.1 for more on this coding convention). The coding convention "centers" the columns in the sense that $\sum_{u=1}^{n} x_{iu} = 0$. Letting $K = [\int_R d\mathbf{x}]^{-1}$, we have

$$J = \frac{nK}{\sigma^2} \int_R E[\widehat{y} - E(\widehat{y}) + E(\widehat{y}) - g]^2 d\mathbf{x}$$

where the quantity $E[\widehat{y}]$ has been added and subtracted. After a couple of straightforward steps we obtain

$$J = \frac{nK}{\sigma^2} \int_R E[\widehat{y} - E(\widehat{y})]^2 d\mathbf{x} + \frac{nK}{\sigma^2} \int_R (E(\widehat{y}) - g)^2 d\mathbf{x}$$

or

$$J = V + B$$

that is, the average or integrated squared error equals the average variance of $\widehat{y}$ (the average prediction variance[5]) V, plus the average squared bias (B). Thus, as pointed out by Box and Draper [21], one could design an experiment to:

1 minimize V. Box and Draper called the resulting designs "all variance" designs;

2 minimize B. Box and Draper called these designs "all bias designs";

3 minimize J.

[3]In this section we will assume, as Box and Draper did, that the region R is a (hyper)sphere. Some of the results, in particular, rotatability, will critically depend on this assumption, as mention later on.
[4]As we will see below, this was one of the points raised by Kiefer in this debate [58, 78].
[5]Minimizing V alone has been called the integrated variance, or IV criterion, in the literature.

One surprising result obtained by these authors is that B can be minimized if the true unknown system is a polynomial model of order d_2 and all we know about this *unknown* polynomial model is its order, d_2, with $d_2 > d_1$. Thus, the parameters of the true model need not be known. In contrast, to minimize J Box and Draper showed that the true parameter values are needed.

We will illustrate these ideas with the case $d_1 = 1$, $d_2 = 2$, and $k = 1$, that is, one controllable factor. This was considered in [21], who actually considered the case $k \geq 1$. The case $d_1 = 2$, $d_2 = 3$ was considered by Box and Draper in [22].

8.2 The Case of a First Order Model ($d_1 = 1$) and Second Order True Process ($d_2 = 2$)**

Consider the case[6] when there is one factor ($k = 1$). In this case, R is both a "spherical region" and a "cuboidal" region of interest. The only factor is coded into $[-1, +1] = R$. The fitted model (using least squares) is

$$\widehat{y} = b_0 + b_1 x \qquad (d_1 = 1)$$

but suppose that the true system is such that

$$E[y] = \beta_0 + \beta_1 x + \beta_{11} x^2 \qquad (d_2 = 2 > d_1 = 1).$$

Since $k = 1$, the design matrix is simply $\boldsymbol{D} = [x_1, \ldots, x_n]'$ where it is assumed the vector adds up to zero due to coding. For the assumed model,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

and

$$J = \frac{nK}{\sigma^2} \left\{ \int_{-1}^{1} (\, Var(b_0) + x^2 \, Var(b_1)) dx \right. \\ \left. + \int_{-1}^{1} (E(\widehat{y}) - \beta_0 - \beta_1 x - \beta_{11} x^2)^2 dx \right\}$$

[6]This section has somewhat more advanced material and may be skipped on a first reading.

where $K = [\int_{-1}^{1} dx]^{-1} = 1/2$. Thus (see Problem 1),

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{n[11]} \end{bmatrix}$$

where $[11] = \sum_{u=1}^{n} x_u^2$ is the second moment of the design (see Section 5.1). This matrix implies that $\text{Var}(b_0) = \sigma^2/n$ and $\text{Var}(b_1) = \sigma^2/(n[11])$. Therefore, the V component of J is

$$\begin{aligned} V &= \frac{nK}{\sigma^2} \int_{-1}^{1} \left(\frac{\sigma^2}{n} + x^2 \frac{\sigma^2}{n[11]} \right) dx \\ &= \frac{1}{2} \int_{-1}^{1} \left(1 + \frac{x^2}{[11]} \right) dx = 1 + \frac{1}{3[11]} \end{aligned}$$

From this it can be seen that to minimize V alone, we should set $[11]$ as large as possible, i.e., we should spread the points of the experiment as much as possible.

To find B in order to minimize it, we need the Alias matrix of the model (see Section 3.6 for a description of the Alias matrix). For the assumed model we have

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

and from the true process we have

$$\boldsymbol{X}_2 = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{bmatrix}.$$

Then, the Alias matrix is therefore

$$\boldsymbol{A} = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2 = \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{n[11]} \end{bmatrix} \begin{bmatrix} n[11] \\ n[111] \end{bmatrix} = \begin{bmatrix} [11] \\ \frac{[111]}{[11]} \end{bmatrix}.$$

Since $E[\widehat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{A}\boldsymbol{\beta}_2$ (see Section 3.6) with $\boldsymbol{\beta}_1 = (\beta_0, \beta_1)'$ and $\boldsymbol{\beta}_2 = \beta_{11}$, we have that

$$E[b_0] = \beta_0 + [11]\beta_{11}$$

and

$$E[b_1] = \beta_1 + \frac{[111]}{[11]}\beta_{11}$$

and therefore

$$E(\widehat{y}) = E(b_0) + E(b_1)x = \beta_0 + \beta_1 x + \beta_{11}\left\{[11] + \frac{[111]}{[11]}x\right\}.$$

Thus, the average squared bias is

$$B = \frac{n\beta_{11}^2}{2\sigma^2}\int_{-1}^{1}\{[11] + \frac{[111]}{[11]}x - x^2\}^2 dx.$$

After some algebra we get

$$B = \frac{n\beta_{11}^2}{\sigma^2}\left\{[11]^2 - \frac{2}{3}[11] + \frac{1}{5} + \frac{[111]^2}{3[11]^2}\right\}. \tag{8.2}$$

Clearly, B is minimized with respect to $[111]$ by making $[111] = 0$. This implies a symmetric design. If $[111] = 0$, the average squared deviation is

$$J = V + B = 1 + \frac{1}{3[11]} + \frac{n\beta_{11}^2}{\sigma^2}\left\{([11] - 1/3)^2 + \frac{4}{45}\right\}$$

Thus, the value of $[11]$ that minimizes J depends on β_{11}. Thus, we need to know a true model parameter to minimize J.

Based on this development, Box and Draper [21] pointed out the following:

1 If B is negligible, one can minimize V by spreading the design points as much as possible making $[11]$ as large as possible. This tends to happen in all "alphabetic" optimal designs which look at functionals of the covariance of the fitted model parameters;

2 To minimize B, make $[111] = 0$ (a symmetric design, recall that the third moment is a measure of skewness) and from (8.2) choose $[11] = 1/3$, which implies *points will be less spread than in an "all variance" design*;

3 A sensitivity analysis in the unknown quantity $\sqrt{n}\beta_{11}/\sigma$ shows that the "all bias" optimal value of $[11] = 1/3$ gives values of J that are (typo) very close to their optimal (minimum) value had the parameter β_{11} been known. Unless V is much larger than B, an all bias design should be chosen (see Table 8.1).

This result generalizes for any number of factors. The most general result stated in [21] is the following.

Table 8.1. Sensitivity analysis on the true unknown standardized parameter $\sqrt{n}\beta_{11}/\sigma$ for the case $k = 1$, $d_1 = 1$, $d_2 = 2$. Table gives the optimal value of $[11]$ if this parameter were known; the corresponding minimum J, and the J value that one gets when using the "all variance" and "all bias" solutions instead ($[11] = 1$ and $[11] = 1/3$) (Source: [117])

$\sqrt{n}\beta_{11}/\sigma$	V/B	$[11]$	J (opt.)	$J([11] = 1.0)$	$J([11] = 1/3)$
0	∞	1.0	1.333	1.333	2.0
4.5	1	0.388	3.718	11.521	3.798
7.0	0.4	0.359	6.316	27.465	6.355

All Bias Result. *B is minimized when all design moments of order up to d_1+d_2 equal the moments of a uniform distribution over R.*

This is a sufficient condition to minimize B, but not necessary, as there may be other designs without this property that could minimize B. The necessary and sufficient condition that must be met to minimize B is

$$\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12} = \boldsymbol{\mu}_{11}^{-1}\mu_{12}$$

where

- $\boldsymbol{M}_{11} = (\boldsymbol{X}_1'\boldsymbol{X}_1)/n$ is the design matrix of moments up to order $2d_1$;
- $\boldsymbol{M}_{12} = (\boldsymbol{X}_1'\boldsymbol{X}_2)/n$ is the design matrix of moments of orders $d_1 + 1$ to $d_1 + d_2$;
- $\boldsymbol{\mu}_{11} = K\int_R \mathbf{x}_1\mathbf{x}_1' d\mathbf{x}$ is the region matrix of moments up to order $2d_1$, and
- $\boldsymbol{\mu}_{12} = K\int_R \mathbf{x}_1\mathbf{x}_2' d\mathbf{x}$ is the region matrix of moments of orders $d_1 + 1$ to $d_1 + d_2$

(here, $\mathbf{x}_i$ is a vector of controllable factors in model form that corresponds with the columns in matrix $\boldsymbol{X}_i$). Thus, one way, but not the only way, when this condition is achieved is to make $\boldsymbol{M}_{11} = \boldsymbol{\mu}_{11}$ and $\boldsymbol{M}_{12} = \boldsymbol{\mu}_{12}$.

A few specific corollaries of this general result are as follows.

Corollary 1. (Box and Draper [21].) If $d_2 = 2$ and we fit a model of order $d_1 = 1$ in k factors then B is minimized with a first order orthogonal design with $[ii] = 1/(k + 2)$, that is, well inside R.

Corollary 2. (Box and Draper, [22].) If $d_2 = 3$ and $d_1 = 2$, then B is minimized when moments up to order 5 equal those of a uniform distribution

over R. In particular, if R is a k-dimensional sphere, the uniform distribution is over such a sphere, and the design that minimizes B is a *rotatable* design (see Chapter 5).

In all of these cases, these authors conclude that the "all bias" designs are close to those obtained from minimizing the averaged squared error J. "If simplification is to be made in the design problem, it might be better to ignore the effects of sampling variation rather than those of bias" [20].

8.3 The Optimal Design Argument

The results in the previous section have been well disseminated among the applied statistics community. Since this book is tailored to such an audience, it is important that the contrary position be also explained.

In [78], Kiefer begins by indicating that the cited conclusion by Box and Draper "has been accepted with a generality which it does not deserve, and which I doubt that Box and Draper would agree". The reasons for this are both matter of principle and matters of understanding the assumptions behind the Box and Draper analysis. With respect to the former, Kiefer indicates that "if fitting a quadratic, even knowing that this is incorrect, the statistician presumably believes B is relative small; otherwise he would fit a cubic or at least let the data determine what type of curve he should fit" [78]. This seems to be confirmed by Vining and Myers [154] who mention that "one should not be overly concerned with model misspecification if the amount of bias is of a magnitude that results in a large power of detection by the standard lack of fit test".

With respect to the underlying assumptions, Kiefer points out that the designs recommended by Box and Draper depend on the choice of a uniform probability distribution over R, where "there is nothing particularly compelling" about it [78, 58]. Different distributions over R will yield different designs. To this, however, Box and Draper [20] replied, indicating that if the uniform distribution is replaced by a weighting function that has spherical contours, with highest weight in the center (e.g., symmetric multivariate normal) then the "all bias solution" will still call for a rotatable design for the case $d_1 = 2$, $d_2 = 3$. This, of course, will not be the case for a non-symmetric distribution.

A second interesting comment found in [58] is that the Box and Draper criterion J is an average criterion over R, whereas the appropriateness of such

criterion could be debated, for example, why not minimizing instead the *maximum* of the squared error over R? Kiefer points out that a design that performs well under the J criterion may perform bad with respect to a maximum squared error criterion. Looking at the maximum of J seems to imply (as pointed out by Vining and Myers [154]) that one should consider how well the design performs over every portion of the region of interest. This comment has resulted in the work conducted during the 1990's on variance dispersion graphs (VDG's), see Section 5.6.

An interesting specific result found by Kiefer and his co-workers was to show that for mixture experiments (i.e., experiments defined over a simplex, see Section 5.8), the Box-Draper recommendation of giving bias more consideration when choosing a design is not justified when $d_1 = 2$ and $d_2 = 3$. This is relevant, as optimal design theory (in particular D-optimal designs) has been used more often in mixture than in non-mixture designs.

One last comment about the generality of the Box-Draper recommendations, which can be seen in action from the above result by Kiefer on mixture experiments, is that it depends on the form of the region R. If R is a cuboidal region, then a rotatable design does not minimize B. This was noted by Box and Draper [20]. As mentioned by Myers and Montgomery [117], rotatability is of interest mainly for a spherical region of interest.

8.4 Conclusions

The debate between bias and variance when deciding on an experimental design, and in general, between the applied school of experimental design and the optimal design school, still reappears in the literature every now and then. A consensus among most authors in industrial practice, however, appears to be that D-optimal designs are extremely useful when the design region is highly constrained, a situation common in mixture experiments, and when there is a good level of confidence about the assumed model. Even in non"mechanistic" situations where polynomial models are used as empirical local approximations, D optimal designs have been found useful, provided the region of the design is not too large compared to the area of operability of a process. A case in point is Robust Parameter Design (see next chapter), where a model that incorporates a quadratic polynomial in the controllable factors and interactions in the noise and control factors (see 9.1) has been found useful in many

practical experiments. Optimal design theory is also very useful for the design of split-plot designs [64]. Another area where D-optimal designs have been found to be very useful is to determine additional experimental runs that augment a given design and resolve effects that are confounded in 2^{k-p} fractional factorial designs.

Experimental design is a multicriteria decision making problem, and it is up to the experimenter to decide what are are his/her priorities and, hence, his/her design criteria and constraints. There are several optimality criteria one could select (see Chapter 5), to which one could add other practical considerations. This is a point emphasized by Box and Draper [20], who give a list of practical considerations when choosing a design. These considerations include the ability to *detect* bias of a specific kind (e.g., ability to perform a curvature test) or of a general kind (e.g., ability to perform lack of fit tests). Also, minimizing the number of runs is almost always an important consideration in practice. Symmetry, balance, number of levels per factor, geometry of the region R, etc., are all constraints that can easily be accommodated by properly chosen constraints in optimization algorithms for experimental design. We refer readers to Section 5.7.8 where good practices are given when using D-optimal designs.

As mentioned by Kiefer, *the ability to detect bias is, from a practical perspective, a very important feature of a good design*. This makes more sense than designing with respect to a *specific* "primary" model M_1 but to desire protection against a second *specific* "secondary" model M_2. Some recent work has suggested to find a design for a family of potential models using a weighted D-optimality criteria [68] but it is not clear how to *a priori* assign the weights. A way to do this is to use Bayesian techniques, in which a first design is run and used to evaluate the posterior probability of each of the models entertained. These probabilities can then be used to design a second experimental design that protects against each potential model according to the posterior probabilities of each model. See Section 12.6 for a description of how to compute posterior probabilities for different models given an initial design. The overall approach may be costly in terms of number of runs. A related body of work is the extensive literature on Bayesian D-optimal designs. In Section 12.9 we describe a simple Bayesian modification of D-optimal designs developed by DuMouchel and Jones [50] that protects against model misspecification.

A final note is that if bias is an overwhelming preoccupation, the experimental design problem approaches that of a "computer experiment" performed with a deterministic simulation or engineering program. In this case, specific designs have been proposed to take into consideration bias only concerns (see Section 14.2 for an introduction).

Once we have reviewed the bias vs. variance debate, we move on to discuss an important class of process optimization problems in the next part of the book.

8.5 Problems

1 Find the $\boldsymbol{X}'\boldsymbol{X}$ matrix and its inverse for the case $d_1 = 1$, $d_2 = 2$, $k = 1$.

2 Suppose that in a response surface study we have fitted the one factor model $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$, but the true response is given by $E[y] = \beta_0 + \beta_1 X + \beta_{11} X^2$. The region of interest is $-1 \leq X \leq 1$. In this RSM study, the purpose is to use the predictor of the *slope* of the response. Find the integrated mean square error $J = V + B$ for the slope of the response. (Hint: the true slope is $\beta_1 + 2\beta_{11} X$ and the estimated slope is $\widehat{\beta}_1$). Based on your results, what guidelines can be given about a design that minimizes V or B for the slope?

3 Consider the case where we have $k = 3$ factors, and we want to fit a first-order model. It is desired to minimize the integrated square *bias* (B) of the estimated response. It is desired to have protection against bias due to a second order model. The region of interest is the cuboidal region $-1 \leq X_i \leq 1$ for $i = 1, 2, 3$.

a) Find moment conditions that minimize B.

b) Suppose that in a) above we wish to use a full factorial with 4 center run replicates. What is the factorial distance (from the design center) that minimizes B? Use the results developed in part a).

PART IV

ROBUST PARAMETER DESIGN AND ROBUST OPTIMIZATION

Chapter 9

ROBUST PARAMETER DESIGN

The secret of my success in competitions was to be prepared against everything.

—Al Oerter[1]

Just as success in competitive sports, finding process settings and product design parameters that are "prepared" against any eventuality or uncertainty is also the basic idea followed in industry to obtain *robust* processes and products. In this chapter we consider robustness with respect to variation in uncontrollable factors, also called noise factors, a problem that has received the name "Robust Parameter Design" (**RPD**), a term coined by Taguchi. Genichi Taguchi [149], a textile engineer with a training in statistics, introduced a series of innovative ideas in designed experiments and process optimization which have had strong influence in the way we look at process optimization today. Some of Taguchi's ideas and concepts have been criticized by several authors, mainly in the USA. This chapter first discusses the main ideas behind Taguchi's approach to the RPD problem. In later sections, we describe how the same goals and ideas introduced by Taguchi can be approached using response surface techniques, including techniques developed relatively recently in answer to the controversy created by Taguchi in quality control and Applied Statistics circles.

One basic idea in Taguchi's approach to process optimization is that, while a process is in operation, not all factors are controllable. There is a need in practice to develop on-line process control mechanisms to compensate against such

[1] American athlete, 4 times consecutive discus olympic gold medal (1956, 1960, 1964, and 1968).

"disturbances", or, as Taguchi called them, "noise factors". This is evidently not a new idea in engineering. What was a very useful idea was to propose that in experiments in industry, under carefully controlled conditions, many if not all of these noise factors can be manipulated and hence, used in a designed experiment. One should then vary these noise factors, which are uncontrollable and noisy while the process operates, and see which system configuration in the controllable factors is the most insensitive, or robust, to the noise factor variation.

Taguchi distinguishes System design, Parameter design, and Tolerance design. System design is a general approach to designing a process that includes defining its objectives and goals. In industry this concerns developing the concept of a particular product and its functionality based on perceived market needs. Parameter design involves defining responses of interest to the goals of the system and optimize them with respect to their mean and variation. Tolerance design corresponds to controlling the variables that have been optimized in the Parameter design step by controlling the factors that affect them. Roughly speaking Parameter Design and Tolerance Design correspond to Process Optimization and Process Control, as described in Chapter 1, with optimization preceding control.

There is no doubt that Taguchi took the Robust Parameter Design idea from standard engineering practice, and this explains in part its popularity in practice. For example, in the design of electronic circuits, engineers have for many years conducted simulations in which uncontrollable factors are varied randomly and the most resistent or insensitive design is chosen. The merit of Taguchi was to introduce this concept in experimental design, and to propose a method, which, although flawed in several ways, generated considerable thinking and newer techniques to solve such "robust parameter design" problems.

The idea of robustness has been around in the Statistics literature for many years as well. Robustness typically has referred to models and methods and the conclusions based on such models that are insensitive to variation in the assumptions. However, robustness with respect to variations in uncontrollable noise factors (both occurring during manufacturing and while the product is being used by end customers) was a new concept introduced by Taguchi[2].

[2]There is a lot of controversy about what are the real contributions by Taguchi, and whether even his idea on robustness has been proposed by earlier writers than him or not. G. Box points out that the idea of robustness

A *robust design* of a product or process is one in which two different types of factors are varied: controllable factors and noise factors. Noise factors cannot be changed or varied during operation of the process or while the product is used in the market place. For example, instances of noise factors in a discrete manufacturing process are ambient temperature and humidity, two factors that can certainly be controlled by placing a machine tool in a "clean room" in order to conduct the experiments (such a facility will in general be extremely expensive for regular production in traditional metal machining manufacturing). Noise factors might also be user-dependent, or relate to variables that occur in the marketplace, once a product is used. An instance of these is in the design of car components. The driving habits or terrain conditions a car or car component will experience once bought by the end customer are certainly uncontrollable for the manufacturer, but they can be varied or simulated in an experiment for the purposes of designing a robust vehicle.

9.1 Optimization Approaches Suggested by Taguchi

Taguchi further distinguishes between factors that have a location effect, i.e., factors that change the mean of the quality characteristic or response, and factors that have a dispersion effect, i.e., factors that affect the variance of the process.

An important message here is that we should always consider the distribution of the quality characteristic when optimizing a process. Taguchi suggested to look at the first two moments of the distribution of a random variable in a peculiar and quite flawed way using signal to noise ratios, which we describe next. Before doing that, let us comment on the experimental strategy Taguchi recommended for robustness. These are **crossed arrays**, experiments that results from the product of two DOE, one in which controllable factors are varied and

of a production process to environmental variation was probably first mentioned by Michaels in a 1964 paper [100], but as pointed out by Box, the paper did not get much attention [25]. The paper by Michaels also was a first reference on the potential of split plot designs to model the relation between products and environmental variation (see Section 9.3 below), with detergent manufacturing used as an example. This anticipated much of the recent work on split plot designs for use in RPD. However, the section where Michaels covers these ideas is rather short (half a page) and it was probably because of this that his paper went mostly unnoticed. Apart of the fairness of Taguchi's critics with respect to the actual *techniques* used by Taguchi, for the most part quite fair, critique of Taguchi's *idea* of robustness to noise factors have followed the famous adage that says that when someone proposes something new that others dislike for some reason, first it is said that it is wrong, then it is said it is unnecessary and trivial, and finally, it is said that it is important but someone else did it before. Criticisms of Taguchi's robustness idea seem to have gone through all 3 stages.

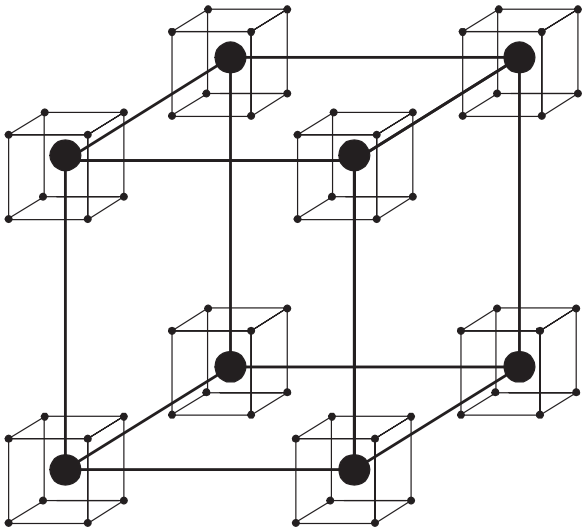


Figure 9.1. An instance of a crossed array experiment. Three control factors varied in the inner array according to a 2^3 design (larger cube) *crossed* with a 2^3 design used to vary the three noise factors in the outer array (smaller cubes)

one in which noise factors are varied (see Figure 9.1). In the figure, eight runs of the *inner array*, the 2^3 DOE used to vary the control factors, are *crossed* with the eight runs of a 2^3 *outer array* used to vary the noise factors. The resulting DOE has 64 runs (the smaller dots on the figure), if run unreplicated. We will comment about the properties of these designs below. For the time being let us note that these designs may result in a large number of runs due to "crossing". Because of this, the Taguchi orthogonal designs that are suggested are frequently fractional factorial designs of low resolution, but this brings other type of problems. The approach is intuitive, however, in the sense of "simulating" variation in the uncontrollable factors at different locations in the space of the controllable factors (inner array).

Assuming a crossed array has been run, Taguchi suggested to consider three types of problems, which correspond to minimization, maximization, and the case the response has a target value:

1 "Smaller the better". Here it is suggested to select the solution as the factor combination in the inner array that

$$\max\ SN_S = -10 \log \sum_{i=1}^{n_0} \frac{y_i^2}{n_0}$$

where n_0 is the number of runs in the outer array.

2 "Larger the better". Here we pick as solution the inner array point that

$$\max\ SN_L = -10 \log \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{1}{y_i^2}.$$

3 "Target is best". Here Taguchi proposed a 2-step approach, which probably has not been criticized *enough*. He suggests two cases:

(a) μ_y (mean of y) is not related to σ_y. In this case the 2 step approach is:

i Select *some* control factors[3] that

$$\max\ SN_{T1} = -10 \log s^2$$

where s^2 is the sample variance taken over the outer array.

ii Select some other inner array factor (not varied before) to make

$$\overline{y} \approx T$$

where T is the target of the quality characteristic.

(b) If σ_y is proportional to μ_y, a case likely to occur in practice, then

i Select some control factors to

$$\max\ SN_{T2} = -10 \log \frac{\overline{y}^2}{s^2}$$

where $\overline{y}$ and s^2 are the average and variance of the outer array observations.

ii Select some other control factor, not varied before, to make

$$\overline{y} \approx T.$$

Since the standard deviation is proportional to the mean, the idea is that the controllable factors will change the mean but will not change the ratio $\frac{\overline{y}^2}{s^2}$ much. Note how this is the only truly "signal to noise" ratio.

[3]Ross [138] suggests to select all except one.

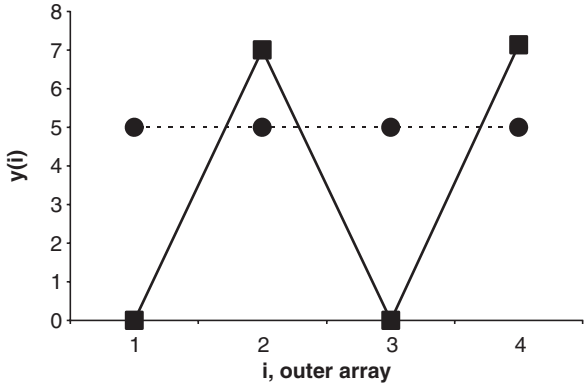


Figure 9.2. Two very different data sets of size $n_0 = 4$ that result in the same SN_S value of -13.97

In all cases, the logarithm and the negative are optimized so that all SN statistics are always *maximized*, regardless of the objective for the response (the factor 10 is completely superfluous).

Signal to noise ratios have been criticized for several good reasons (for a review, see the Technometrics issue edited by V. Nair [118]). In particular:

- A data set with no outer array variability and one with considerable outer array variability may result in the same SN statistic (Figure 9.2). This was noted by Box [18]. The implication is that no Robust Parameter Design can be done;
- Factors may have both location and dispersion effects, so the proposed 2-step approach may simply be infeasible to do in practice.
- The suggested 2-step approach seems more like wishful thinking than like a general optimization strategy that solves trade-offs in as much as possible, as formal multi-objective and non-linear optimization techniques do[4].

The problems related to the crossed arrays will be discussed further in the next section.

9.2 Control $\times$ Noise Interaction Effects

Some authors have suggested looking at 2-factor interaction plots for interactions that occur between a controllable (C) factor and a noise (N)

[4]OR students with a strong optimization background always find this naive 2-step approach quite amazing.

factor. The importance of these interactions is that it is assumed the non-homogeneous variance found over the controllable factor space is due to the presence of noise factors. If there is a significant $C \times N$ interaction, then one can choose the value of the C factor to *neutralize* the variability transmitted by the N factor. Interaction plots are useful for this purpose.

For instance, suppose we have conducted an experiment with 3 control factors (A, B, and C) and one noise factor (D) and let us suppose the AD interaction turns out to be significant. We wish to maximize the response y. Suppose the AD interaction plot looks as in Figure 9.3. We need to imagine what would happen if factor D is left to vary randomly within its low and high value, uncontrolled, once the process is released to manufacturing or when the product is sold and used in the market place. We can set factor A "in the factory". If we set A to its "low" value, this will result in wild variations of the response. If we set A to its high value, we can see the response will vary little with variations of the noise factor. This setting, by a happy coincidence, also results in higher response settings. Hence, setting $A = +1$ is a solution to this maximization problem that is robust with respect to variations in the noise factor D.

In this case we were lucky. If the problem had been instead to minimize the response, then we would have not known how to proceed to achieve a robust minimizing solution.

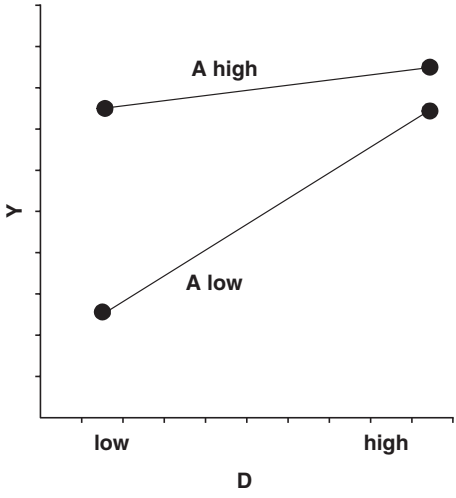


Figure 9.3. A hypothetical control $\times$ noise interaction: A is controllable and D is a noise factor

A key point to notice is that to reduce the sensitivity of the response to variations in the noise factors, we select settings for which the *slope* of the response with respect to the noise factors (as displayed, e.g., on the control $\times$ noise interaction plot) is *flattest*. This is exploited in more formal mathematical programming approaches to solving robust parameter design problems, as shown later in this chapter. We first take a closer look at experimental strategies that have been proposed for RPD.

9.3 Experimental Designs Used in RPD

Crossed arrays of the type promoted by Taguchi can be shown to always provide $C \times N$ interactions that are not aliased with main effects or other two factor interactions [142]. As mentioned earlier, these interactions are vital in any RPD study, because they allow us to neutralize the effect of the noise factors – which are uncontrollable – by manipulating the controllable factors. If a process has no significant $C \times N$ interactions or these interactions are aliased and cannot be estimated clearly, the RPD problem cannot be solved.

However, although $C \times N$ interaction are estimable if using crossed arrays, the designs themselves usually require a large number of runs. To reduce the number of runs, fractional factorials are run in the control (inner) array. If this is done, the usually important $C \times C$ interactions will not be estimable. A simple example used to demonstrate this problem (see Shoemaker et al. [142]) is an experiment with three control factors (x_1, x_2, x_3) and 3 noise factors (z_1, z_2, z_3). The inner array is a 2^{3-1} with defining relation $I = x_1x_2x_3$ and is crossed with the outer array, a 2^{3-1} design with defining relation $I = z_1z_2z_3$. The 16 run experiment then has a complete defining relation[5] equal to

$$I = x_1x_2x_3 = z_1z_2z_3 = x_1x_2x_3z_1z_2z_3.$$

In this design:

- $C \times C$ interactions are aliased with main effects, so all main effects are estimable only if the $C \times C$ and the $N \times N$ are not existent;
- the noise factor main effects are aliased with a $N \times N$ interaction (i.e., resolution = III among the noise factors);

[5]The complete defining relation of a crossed array can be found by taking all the possible products of the generators in each of the inner and outer array and adding the inner/outer array generators themselves.

- The crucial $C \times N$ interactions are better estimated as they are aliased only with a 4-word chain of 3 and 4 letter interactions.

One could use a 2^3 as an inner array instead. This will allow to estimate all $C \times C$ interactions and C main effects clear of any aliasing. If a 2^{3-1} is still used as outer array, the noise main effects will be aliased with $N \times N$ interactions, as before. Each $C \times N$ interactions will be aliased only with one 3-letter interaction. However, the design is now 32 runs large.

An alternative (see Shoemaker et al. [142]) is to use a single or **combined** array consisting of a 2^{6-2} design with 16 runs and defining relation:

$$I = x_1x_2x_3z_1 = z_1z_2z_3 = x_1x_2x_3z_2z_3.$$

This design has **mixed resolution**:

- It also has resolution III among the noise factors;
- Each $C \times N$ interaction is aliased with a $C \times C$ interaction and 2 other 3 and 4 letter interactions (i.e., resolution IV in the control factors).

Thus, we have gained that the control main effects are *not* aliased with 2-factor interactions. This is a clear advantage.

The previous example and similar examples of mixed resolution designs in the literature do not provide a systematic way of finding designs with good alias properties. A more systematic approach can be achieved by fixing the model form. A useful model for RPD when the model is quadratic in the controllable factors was proposed by Box and Jones [27] (see also [113]). Suppose there are k controllable factors $\mathbf{x}' = [x_1, x_2, \ldots, x_k]$ and r noise variables $\mathbf{z}' = [z_1, z_2, \ldots, z_r]$. Then the model is:

$$Y_i(\mathbf{x}_i, \mathbf{z}_i) = \beta_0 + \mathbf{x'}_i\boldsymbol{\beta} + \mathbf{x}_i'\mathbf{B}\mathbf{x}_i + \mathbf{z}_i'\boldsymbol{\gamma} + \mathbf{x'}_i\boldsymbol{\Delta}\mathbf{z}_i + \epsilon_i \quad \forall i = 1, \ldots, n \quad (9.1)$$

where:

- $Y_i(\mathbf{x}_i, \mathbf{z}_i)$ is the value observed for the response in the i^{th} experiment given a fixed value of the noise variables, $\mathbf{z}_i$, and the corresponding values of the controllable factors $\mathbf{x}_i$;
- $\{\epsilon_i\}$ is a set of normal i.i.d. random variables with zero mean and variance σ_ϵ^2;

- β_0, $\boldsymbol{\beta}$, $\mathbf{B}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$ are the model parameters. In particular, $\boldsymbol{\gamma}$ is a $r \times 1$ vector with the noise factors main effects and $\boldsymbol{\Delta}$ is a $k \times r$ matrix containing all the control $\times$ noise interactions.

Borkowski and Lucas [13] (see also [89]) proposed a family of **composite mixed resolution (CMR)** designs. These designs consist of:

- A 2-level fractional factorial design with:
  - resolution V among the control factors;
  - at least resolution III among the noise factors;
  - no $C \times N$ interaction aliased with main effects or other 2-factor interaction;
- $2k$ axial runs in the control factors. Borkowski and Lucas used $\alpha = 1$;
- N_0 center points.

The conditions imposed on the factorial part do not guarantee uniqueness. Therefore, the additional criterion of minimum aberration was used by Borkowski and Lucas [13] to find specific CMR designs. Table 9.1 is a sample of some of the CMR designs presented by these authors.

Example. CMR Design. Suppose we feel model (9.1) is adequate for a problem with 3 noise factors and 3 controllable factors. We can then use design 6A (Table 9.1), assigning the first 3 columns to the control factors and the last 3 columns to the noise factors. We add 6 axial points and some center runs to complete the design (Table 9.2). ■.

Table 9.1. Selected Fractions of CMR designs (source: Borkowski and Lucas [13])

Design	$k+r$	k	Fraction	Control Factors	Noise Factors
4A	2	2	2^4	A, B	C, D
5A	5	2–3	2_V^{5-1}	A, B, C	D, E=ABCD
6A	6	2–4	2_{VI}^{6-1}	A, B, C, D	E, F=ABCDE
7A	7	2–3	2_{IV}^{7-2}	A, B, C	D, E, F=ABCE, G=ABCD
8A	8	2	2_{IV}^{8-3}	A, B	C, D, E, F=ABCE

Table 9.2. CMR design based on the fraction 6A for 3 control and 3 noise factors

A	B	C	D	E	F=ABCDE
−1	−1	−1	−1	−1	−1
1	−1	−1	−1	−1	1
−1	1	−1	−1	−1	1
1	1	−1	−1	−1	−1
−1	−1	1	−1	−1	1
1	−1	1	−1	−1	−1
−1	1	1	−1	−1	−1
1	1	1	−1	−1	1
−1	−1	−1	1	−1	1
1	−1	−1	1	−1	−1
−1	1	−1	1	−1	−1
1	1	−1	1	−1	1
−1	−1	1	1	−1	−1
1	−1	1	1	−1	1
−1	1	1	1	−1	1
1	1	1	1	−1	−1
−1	−1	−1	−1	1	1
1	−1	−1	−1	1	−1
−1	1	−1	−1	1	−1
1	1	−1	−1	1	1
−1	−1	1	−1	1	−1
1	−1	1	−1	1	1
−1	1	1	−1	1	1
1	1	1	−1	1	−1
−1	−1	−1	1	1	−1
1	−1	−1	1	1	1
−1	1	−1	1	1	1
1	1	−1	1	1	−1
−1	−1	1	1	1	1
1	−1	1	1	1	−1
−1	1	1	1	1	−1
1	1	1	1	1	1

Table 9.2. (Continued)

-1	0	0	0	0	0
1	0	0	0	0	0
0	-1	0	0	0	0
0	1	0	0	0	0
0	0	-1	0	0	0
0	0	1	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

If one is quite certain that model (9.1) is an adequate representation of the response, then a D-optimal design is certainly a possibility that should be considered.

9.4 Split Plot Designs and Robust Parameter Design

The difficulty with which a factor can be varied from run to run in an experiment may restrict the number of times we change that factor in a DOE. This implies a restriction on the randomization akin to blocking[6]. If randomization restrictions exist, the ANOVA must take them into account. The resulting designs are called **split plot** or *split unit* designs. In this section, we provide an overview of split plot designs and their use in RPD.

As an instance of a split plot design in robust parameter design, consider a car engine testing scenario [10]. A team of engineers wishes to study different configurations of an engine (created according to a 2^{k-p} design) and determine their robustness with respect to noise factors that are varied according to a 2^{q-r} design. The noise factors are easier to vary, e.g., using different gasoline types. However, to start an experimental run, each engine prototype needs to be placed on a test stand. Considerable time is spent in setting up each test. Hence, the engineers would prefer testing each engine prototype whose configuration makes up a "point" in the inner array, at every combination of the noise factor

[6]Although as we will see shortly, this differs from blocking in that a factor is introduced between blocks.

treatments, which are then tested in random order. Running the experiment in this way (a split-plot) the engineering team will save setup time and cost.

Notice, however that this experiment:

1 is not completely randomized. Randomization occurs only at the outer array (noise factors) level. This needs to be accounted for in the ANOVA;

2 has effects that will be estimated with different precisions. Intuitively, the effects associated with noise factors will be estimated more precisely than those related to control factors, since the different engines act as blocks. This means that noise factors are contrasted within an experimental unit that is more homogeneous (the *same* engine in this case). In contrast, the different configurations (type of carburetor, filters, etc.) are applied to a more variable kind of experimental run, namely, *different* engines.

A split-plot experiment can be thought of as a blocked experiment in which a set of factors are introduced and varied from block to block (and hence, their effects will be confounded with blocks).

Split-plot designs originated in Agriculture experiments. A large plot of land may be an adequate experimental unit to apply a factor such as different pesticides (e.g., pesticide application by airplane is inexpensive but has to be done over large plots). However, the large plots or **whole plots** of land are unnecessary or inappropriate to apply other factors, such as type of seed, which can be applied easily over small plots. Thus the whole plots are divided or split in smaller **subplots** (or split-plots) of land, to which the easy to vary factors can be applied. If our main interest is the effects of the seeds (subplot factor) rather than the effect of the pesticides (whole plot factor), then running the experiment as a split plot will not only save money to the experimenter with respect to a completely randomized experiment, but it will give also more precise estimators of the effects associated to the subplots. The "price to pay", is less precision in the whole plot factor effects.

Because of the two different "sizes" of experimental units, whole plot and split plot factor effects need to be compared against different error terms to determine their significance. Typically, variation between large units (whole plots) is larger than variation between small units (split plots), because the latter are more homogeneous than the former. Thus, the significance of the factor that varies from whole plot to whole plot should be determined by comparison

against the whole plot experimental error. Likewise, the significance of the subplot factor effects should be determined by comparison against the subplot experimental error.

In industrial experiments, split plotting arises due to **hard to vary factors**. These factors sometimes arise, as in the engine testing mentioned before, because of setup times in preparing a test of a prototype. In chemical industries, split plotting arises because a large batch of a mix is first produced by varying the factors in a chemical formulation which can then be divided in smaller batches for further experimentation with respect to other factors.

We now describe how to analyze split plot designs. We will start with 2-factor experiments.

9.4.1 Split-plot Designs with 2 Factors

Suppose an experimenter has two factors of interest in addition to a nuisance factor (say, the day in which the experiments are conducted), the latter which she wishes to block. The experiment was run following the split-plot design of Figure 9.4.

A model useful to explain the response that is collected from this experiment is as follows:

$$y_{ijk} = \mu + b_i + p_j + \varepsilon'_{ij} + q_k + (pq)_{jk} + \varepsilon_{ijk}, \quad \begin{cases} i = 1, \ldots, B \\ j = 1, \ldots, P \\ k = 1, \ldots, Q \end{cases} \tag{9.2}$$

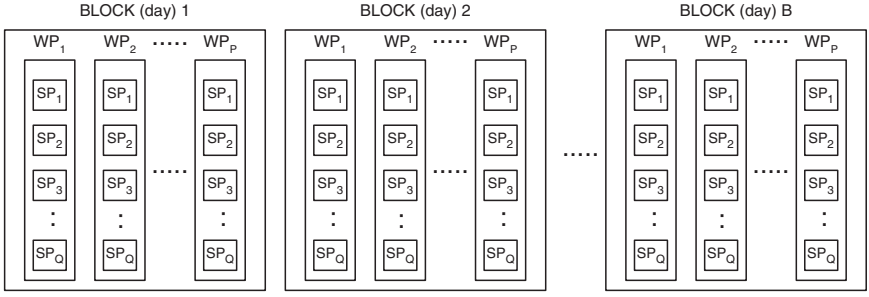


Figure 9.4. A two factor split-plot arrangement where there is a nuisance factor (days) that needs to be blocked

Here, b_i is the block i effect, p_j is the effect of whole plot j, q_k is the effect of subplot k and $(pq)_{jk}$ is the interaction effect of the j level of the whole plot factor and the k level of the subplot factor.

The factor applied over the whole plots has P levels (one per "whole plot") and a second factor is varied at random within each whole plot across Q levels, one per subplot. The most important feature of model (9.2) is that *there are two error terms*, ε'_{ij}, which is the whole plot to whole plot error, and the subplot to subplot error, ε_{ijk}.

The sums of squares for all terms are computed in the same way as a completely randomized three factor experiment (see Appendix B for notation):

$$SS_{\text{WP factor}} = \frac{1}{BQ}\sum_{j=1}^{P} y_{\bullet j\bullet}^2 - \frac{y_{\bullet\bullet\bullet}^2}{BPQ}$$

$$SS_{\text{SP factor}} = \frac{1}{BP}\sum_{k=1}^{Q} y_{\bullet\bullet k}^2 - \frac{y_{\bullet\bullet\bullet}^2}{BPQ}$$

$$SS_{\text{Blocks}} = \frac{1}{PQ}\sum_{i=1}^{B} y_{i\bullet\bullet}^2 - \frac{y_{\bullet\bullet\bullet}^2}{BPQ}$$

$$SS_{\text{Blocks*WP}} = \frac{1}{Q}\sum_{i=1}^{B}\sum_{j=1}^{P} y_{ij\bullet}^2 - \frac{y_{\bullet\bullet\bullet}^2}{BPQ} - SS_{\text{Blocks}} - SS_{\text{WP}}$$

$$SS_{\text{WP*SP}} = \frac{1}{B}\sum_{j=1}^{P}\sum_{k=1}^{Q} y_{\bullet jk}^2 - \frac{y_{\bullet\bullet\bullet}^2}{BPQ} - SS_{\text{WP}} - SS_{\text{SP}}$$

Assuming the blocks are random effects, the Blocks $\times$ Whole Plots interaction sum of squares is the whole plot error sum of squares. The sum of squares for the subplot error is obtained by subtraction. The corresponding degrees of freedom are shown in Table 9.3.

Example. A 2-factor Split Plot design. Consider the engine testing scenario described earlier. Suppose that 3 different engines are going to be tested for emissions of CO, where the response is the PPM of CO in the exhaust gases. It is of interest to test four gasoline additives (A,B,C,D) to determine a reduction in emissions. Since setting up an engine for testing requires considerable setup time, the engineers decided to run the experiment as a Split Plot by testing the four gasoline additives in each engine. Three replicates were taken, and the experimental layout and observed responses are shown in Table 9.4. The corresponding ANOVA, obtained using Minitab's General Linear Model option is as follows:

Table 9.3. ANOVA for 2-factor Split-plot design with random blocks

Source	**d.o.f**
Blocks	$(B-1)$
WP main effect	$(P-1)$
WP error	$(B-1)(P-1)$
SP main effect	$(Q-1)$
WP $\times$ SP interaction	$(P-1)(Q-1)$
SP error	By subtraction: $(B-1)P(Q-1)$
Total	$BPQ-1$

Table 9.4. Experimental layout and data, 2-factor Split plot engine prototype experiment

	Replicate 1			**Replicate 2**			**Replicate 3**		
Additive	1	2	3	1	2	3	1	2	3
A	1386	3206	2147	2752	1417	2126	705	2103	1544
B	1276	2930	2316	2903	1823	1643	503	1921	1557
C	2200	2690	1680	3245	1207	1870	1439	1464	1196
D	1333	2384	1714	3013	1432	1911	565	1492	1233

```
Source             DF    Seq SS   Adj SS   Adj MS      F      P
Replicate           2   5099374  5099374  2549687   1.08  0.421
Engine              2    486149   486149   243075   0.10  0.904
Replicate*Engine    4   9435638  9435638  2358909  61.63  0.000
Additive            3    350917   350917   116972   3.06  0.055
Engine*Additive     6   1620242  1620242   270040   7.05  0.001
Error              18    689005   689005    38278
Total              35  17681326

Variance Components, using Adjusted SS

                  Estimated
Source                Value
Replicate             15898
Replicate*Engine     580158
Error                 38278
```

The Replicates and Engines terms are tested against the Replicates × Engines interaction, which is a random effect and acts as the whole plot error. The Additive and Engine × Additive terms are tested against the "error" term on the table which corresponds to the subplot error, which is obtained by subtraction. As it can be seen, the subplot factor (Gasoline additive) is significant, and so is the Engine × Additive (WP × SP) interaction. This is good news for the engineers, because it means they can select a gasoline additive to reduce the emissions. This can be observed from Figure 9.5, from which it can be concluded that additive type D reduces more consistently the emissions compared to the other additives. Finally, notice that the variance components estimates yield $\widehat{\sigma_{\text{WP}}} = 761$ (from the Replicate × Engine term) and $\widehat{\sigma_{\text{SP}}} = 195$ (from the "error", the subplot error, term). This is as expected: the WP to WP variance is typically larger than the SP to SP variance. ■

Mead [95] has pointed out that the gain in precision in the subplots in a split plot design, over a completely randomized design, is smaller than the precision lost in the whole plot effects. To see this, consider the two-factor split-plot experiment with blocks in Table 9.3. If the experiment were run in a completely randomized way, the mean square against which every effect would be compared for significance testing would be s^2. This is a weighted average of $\text{MS}_{\text{WP error}}$ and $\text{MS}_{\text{SP error}}$ with weights equal the degrees of freedom of these

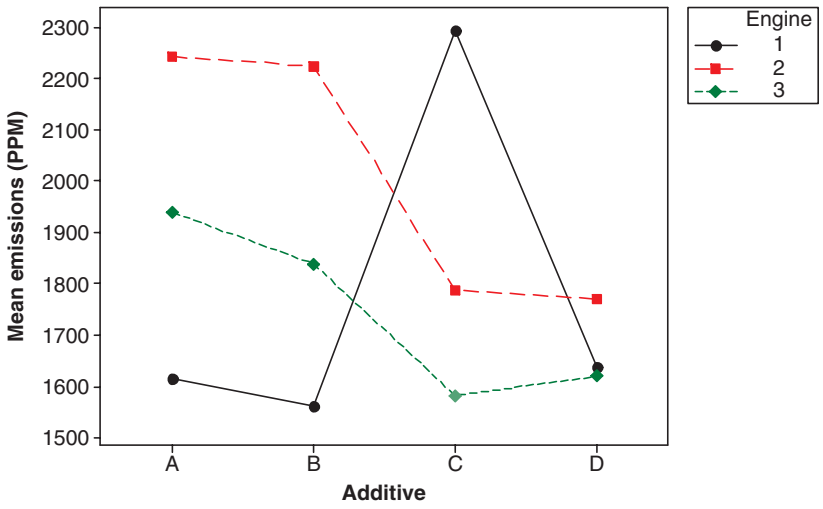


Figure 9.5. Engine type–additive interaction plot, engine testing example

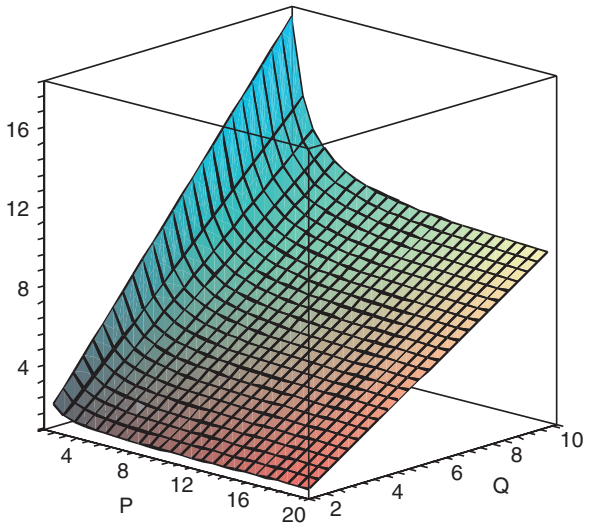


Figure 9.6. $(MS_{\text{WP error}} - S^2)/(s^2 - MS_{\text{SP error}})$ as a function of P and Q

mean squares. It can be shown (see Problem 1) that

$$\frac{\text{MS}_{\text{WP error}} - s^2}{s^2 - \text{MS}_{\text{SP error}}} = \frac{P(Q-1)}{P-1} \tag{9.3}$$

This is a measure of the gain in precision at the SP level at the expense of the loss in precision at the WP level. Notice how the number of blocks does not alter the relative precision of any estimate. In other words, the closer this ratio is to one, the less we lose precision at the WP level relative to the gains at the SP level. It is easy to see (see Figure 9.6) that (9.3) has a minimum at $Q = 2$. Thus, to reduce the lack of precision problem at the WP "stratum"[7] relative to the SP stratum, one should *reduce the number of subplots as much as possible*. Increasing the number of whole plots does help too, as mentioned by some authors [64], but not as dramatically.

9.4.2 Split-plot Designs in 2-level Factorials and their use in RPD

There has been considerable recent work on split plot designs applied to Robust Parameter Design (see e.g., [10, 84, 64]). The Split-Plot nature of any

[7]A terminology used by some authors, e.g., Trinca and Gilmour [151].

experiment arises naturally when one or more factors are hard to vary. Hard to vary factors constitute the whole plot factors, while easier to vary factors constitute subplot factors. In a RPD experiment, in most cases the *control* factors are hard to vary, since they refer to the design of a product or process (e.g., consider the engine testing scenario in the previous example). To save cost, the RPD experiment is not run in a completely randomized (CR) way but as a split-plot. The problem with such approach, as mentioned before, is that the whole plot effects (the control factors in this case and their interactions) will be estimated less precisely than if the experiment were run as a CR experiment. To reduce this problem it has been suggested to increase the number of whole plots. As shown in the previous section for the case of a two factor Split-Plot design, reducing the number of subplots has a stronger effect in improving the precision of the WP effect estimates. We now consider this matter for the case of a general factorial experiment.

Consider the case of a $2^{k-p} \times 2^{q-r}$ crossed array used for RPD. As far as we know, Taguchi did not recommend running crossed arrays in a Split-Plot manner but rather emphasized that they should be run in a completely randomized way. This, actually, is *good* advise: *if an experiment can be run as a completely randomized design, it should not be run as a Split Plot*. Conversely, analyzing a Split Plot experiment as a CRD can lead to conclude erroneously that whole plot effects are significant and subplot effects insignificant. Furthermore, since the C, $C \times C$, and $C \times N$ effects are more important than the N and $N \times N$ effects, only when the hard to vary factors are *noise* factors is when a Split-Plot design (with noise factors varied in the whole plots) is recommended for Robust Parameter Design. This was actually suggested by Michaels [100]. Unfortunately this will not always be possible to do in practice, as control factors can be the harder to vary, like in the prototyping experiment mentioned earlier.

When *control factors* are extremely costly to vary, this will force the experimenter to run a Split Plot design with control factors varied in the whole plots. In such case, the experimenter should try to maximize the number of whole plots (i.e., use the largest 2^{k-p} fraction for the inner array his/her budget allows) and minimize the number of subplots (i.e., use the smallest possible fraction 2^{q-r} for the outer array. Notice how using a very small fraction for the SP stratum enhances the precision (i.e., reduces variance), but at the cost of aliases or bias. This is the ever present variance-bias tradeoff in experimental design also present in Split Plot designs.

Table 9.5. A 2×2^2 crossed experiment for the example

	B	$-$	$+$	$-$	$+$
A	C	$-$	$-$	$+$	$+$
$-$		y_{11}	y_{21}	y_{31}	y_{41}
$+$		y_{12}	y_{22}	y_{32}	y_{42}

Table 9.6. The 2×2^2 Split Plot design with the errors of each observation

A	B	C	i	j	errors
$-$	$-$	$-$	1	1	$\varepsilon'_1 + \varepsilon_{11}$
$-$	$+$	$-$	2	1	$\varepsilon'_1 + \varepsilon_{21}$
$-$	$-$	$+$	3	1	$\varepsilon'_1 + \varepsilon_{31}$
$-$	$+$	$+$	4	1	$\varepsilon'_1 + \varepsilon_{41}$
$+$	$-$	$-$	1	2	$\varepsilon'_2 + \varepsilon_{12}$
$+$	$+$	$-$	2	2	$\varepsilon'_2 + \varepsilon_{22}$
$+$	$-$	$+$	3	2	$\varepsilon'_2 + \varepsilon_{32}$
$+$	$+$	$+$	4	2	$\varepsilon'_2 + \varepsilon_{42}$

To illustrate the difference in precision of WP and SP effect estimates in a factorial experiment, consider an experiment discussed by Bisgaard[8] [10] and shown in Table 9.5.

This is a $2^1 \times 2^2$ Split Plot design with factor A in the whole plot and factors B and C in the subplots. A model for this experiment is

$$y_{ij} = f(\boldsymbol{x}_{ij}) + \varepsilon'_j + \varepsilon_{ij}, \begin{cases} i = 1, 2, 3, 4 \\ j = 1, 2 \end{cases}$$

where $\text{Var}(\varepsilon'_j) = \sigma^2_{\text{WP}}$ and $\text{Var}(\varepsilon_{ij}) = \sigma^2_{\text{SP}}$ and $f(\boldsymbol{x}_{ij})$ denotes the effects of all factors of interest (the important feature of the model is its two-error structure). Table 9.6 shows the design, together with the errors that are part of each of the observed responses if the experiment is conducted as a Split Plot.

From Table 9.6, it can be seen that the variance of the whole plot effect A is

$$\text{Var}(A) = \text{Var}\left(\frac{1}{4}\left(-4\varepsilon'_1 + 4\varepsilon'_2 - \sum_{i=1}^{4}\varepsilon_{i,1} + \sum_{i=1}^{4}\varepsilon_{i2}\right)\right) = 2\sigma^2_{\text{WP}} + \frac{\sigma^2_{\text{SP}}}{2}. \tag{9.4}$$

[8]Who we follow in this illustration.

If instead the experiment had been run as a completely randomized experiment, every observation would contain a WP and a SP error ($\varepsilon''_{ij} = \varepsilon'_i + \varepsilon_{ij}$), so

$$\text{Var}(A) = \text{Var}\left(\frac{1}{4}\left(\sum_{i=1}^{4}\sum_{j=1}^{2}\varepsilon"_{ij}\right)\right) = \frac{1}{16}8(\sigma^2_{\text{WP}} + \sigma^2_{\text{SP}}) \qquad (9.5)$$

which evidently is less than (9.4). So, we get a more precise estimate for the whole plot effects if the experiment is run in a completely randomized way. For a subplot effect, say factor B, we have that

$$\text{Var}(B) = \text{Var}\left(\frac{\sum \pm\varepsilon_{i1} - \sum \pm\varepsilon_{i2}}{4}\right) = \frac{1}{2}\sigma^2_{\text{SP}}$$

since the whole plot error terms ε'_j cancel out. This is evidently smaller than (9.5) which is what we would get had the experimenter ran the design completely randomized.

In general, Bisgaard [10] shows that for a $2^{k-p} \times 2^{q-r}$ crossed array run as a Split Plot, a whole plot factor effect is estimated with variance

$$\frac{4}{N}(2^{q-r}\sigma^2_{\text{WP}} + \sigma^2_{\text{SP}})$$

while a subplot effect is estimated with variance

$$\frac{4}{N}\sigma^2_{\text{SP}}.$$

Notice how to reduce the variance of the WP effects, 2^{q-r}, the number of subplots, should be as small as possible[9].

In the practical analysis of crossed or combined arrays run as Split Plots, all that is necessary is a rule to determine which effects should be compared against the WP error and which effects should be compared against SP error. Bisgaard [10] gives the following simple rule:

Effects involving only WP factors and aliases of these effects are tested by comparing their sum of squares against the WP error. All other effects are tested by comparing against the SP error.

It is furthermore suggested that two separate normal probability plots of the effects should be used, one for WP effects and their aliases, the other for all the other effects, to determine the significant effects.

Example. A crossed factorial array run as a Split Plot. A semiconductor manufacturer wishes to run an experiment in one of their chemical-polishing-

[9]But also as noticed before, the smaller the SP fraction the more aliases or bias.

machines in order to model the non-uniformity of the silicon wafers after polishing as a function of several factors. Eight factors of potential importance are being considered. Difficult to control factors are A=Pad age, B=Incoming non-uniformity, C=Type of Slurry, D=Conditioning of the machine, and E=Routing. These are difficult to change either because they require considerable time (e.g., change the type of polishing pads), or because they refer to properties the wafers acquire before the current processing step (polishing). The three remaining factors are F=polishing time, G=Down pressure, and H=RPM's. These are easier to vary as they are simply entered in the machine in order to process the next run of wafers. Therefore, the engineers decided to run a combined 2_V^{8-2} array ($N = 64$) in Split-Plot form[10]. The design uses generators G=ABCD and H=ABEF and is shown in Table 9.7. Notice how there are $2^5 = 32$ whole plots. In each whole plot only one treatment combination in A, B, C, D, and E is performed. Also, each whole plot contains two subplots, where factors F, G, and H are varied.

In this design, the mean squares for the A, B, C, D, E, AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE effects will be compared against the whole plot error, while the mean squares corresponding to all other effects will be tested against the subplot error. ■

9.4.3 Split Plot Model Estimation

In general, the individual effects from Split Plot designs will not be estimable and hence the main goal in analyzing SP experiments is to conduct the appropriate tests of significance using ANOVA. This was discussed in the previous sections. For designing optimal Split Plot designs it is necessary to know how these models are estimated. Any split plot design and, in general, designs with complicated structure of random and fixed effects can be analyzed using a **linear mixed model** [140, 64] of the form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{X}$ is $n \times p$, $\boldsymbol{\beta}$ is $p \times 1$, $\boldsymbol{Z}$ is $n \times W$ (where W denotes the number of whole plots), $\boldsymbol{\gamma}$ is $W \times 1$ and $\boldsymbol{\varepsilon}$ is $n \times 1$. Here, $\boldsymbol{\varepsilon} \sim (\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$, $\boldsymbol{\gamma} \sim (\boldsymbol{0}, \sigma_y^2 \boldsymbol{I}_W)$ and $\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\gamma}) = \boldsymbol{0}_{W \times n}$. In a Split Plot experiment, the vector $\boldsymbol{\varepsilon}$ models the subplot error, the random vector $\boldsymbol{\gamma}$ models the whole plot errors

[10]If this were a completely randomized design, 64 runs would be considered a large experiment for most people in industry, particularly semiconductor industry. However, the split-plot structure and its associated savings actually allow to run larger experiments

Table 9.7. A 2_V^{8-2} combined array run as a Split plot with A, B, C, D, and E as whole plot factors and F, G, and H as subplot factors

Whole Plot	A	B	C	D	E	F	G=ABCD	H=ABEF	Standard Order	Run Order
1	−1	−1	−1	−1	−1	−1	1	1	1	1
	−1	−1	−1	−1	−1	1	1	−1	17	2
2	1	−1	−1	−1	−1	−1	−1	−1	2	3
	1	−1	−1	−1	−1	1	−1	1	18	4
3	−1	1	−1	−1	−1	−1	−1	−1	3	5
	−1	1	−1	−1	−1	1	−1	1	19	6
4	1	1	−1	−1	−1	−1	1	1	4	7
	1	1	−1	−1	−1	1	1	−1	20	8
5	−1	−1	1	−1	−1	−1	−1	1	5	9
	−1	−1	1	−1	−1	1	−1	−1	21	10
6	1	−1	1	−1	−1	−1	1	−1	6	11
	1	−1	1	−1	−1	1	1	1	22	12
7	−1	1	1	−1	−1	−1	1	−1	7	13
	−1	1	1	−1	−1	1	1	1	23	14
8	1	1	1	−1	−1	−1	−1	1	8	15
	1	1	1	−1	−1	1	−1	−1	24	16
9	−1	−1	−1	1	−1	−1	−1	1	9	17
	−1	−1	−1	1	−1	1	−1	−1	25	18
10	1	−1	−1	1	−1	−1	1	−1	10	19
	1	−1	−1	1	−1	1	1	1	26	20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	−1	1	1	1	1	−1	−1	1	47	61
	−1	1	1	1	1	1	−1	−1	63	62
32	1	1	1	1	1	−1	1	−1	48	63
	1	1	1	1	1	1	1	1	64	64

(one per WP), and $Z_{ij} = 1$ if the ith observation resides in the jth whole plot ($Z_{ij} = 0$ if otherwise). With this, the covariance matrix of the observations is

$$\text{Cov}(\boldsymbol{y}) = \sigma_y^2 \boldsymbol{Z}'\boldsymbol{Z} + \sigma_\varepsilon^2 \boldsymbol{I}_n \equiv \boldsymbol{V}.$$

Because the variances are not constant for all observations, generalized least squares (GLS) needs to be used, giving:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \tag{9.6}$$

with a covariance matrix of the GLS estimators given by

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} \tag{9.7}$$

In practice, the variance components σ_ε^2 and σ_y^2 are not known and must be estimated [84]. The variance estimates are then plugged into (9.6–9.7). The determinant of matrix $\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}$ is used for the D-optimal design of split plot designs and other complex DOEs, see [64]. In some important cases, the GLS estimators reduce to the ordinary least squares (OLS) estimators ($\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$). Two such cases are saturated designs (i.e., designs for which $n = p$) and crossed array designs. Therefore, OLS suffices for designing D-optimal split-plot designs for RPD, provided the experimental design that is desired is a crossed array. See Goos [64] for details.

9.5 Mean and Variance Estimation: A Dual Response Approach to RPD**

Let us return to model (9.1) and its use for RPD[11]. We assume from now on that RPD experiments are run completely randomized.

Once this model is fit[12], we assume that the noise factors will vary either while the process operates or when the product is used according to some distribution with *known* mean equal to zero and known variance, i.e.,

$$\boldsymbol{z} \sim (\mathbf{0}, \boldsymbol{\Sigma}_z).$$

We can then obtain "dual" models for mean and variance of the response by taking expectation and variance with respect to all the random variables in the model, namely, the noise factors $\boldsymbol{z}$ and the uncontrollable error ϵ. This yields functions of the controllable factors $\mathbf{x}$ only, which in principle can be used in a non-linear optimization problem as discussed in Section 4.4. The two functions are:

$$E_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} \tag{9.8}$$

and

$$Var_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}) + \sigma_\epsilon^2 \tag{9.9}$$

Although in principle historical data can be used to estimate $\boldsymbol{\Sigma}_{\mathbf{z}}$, the possibility of errors in this estimate indicates the need for further research on this aspect of the problem, which thus far has not been investigated.

Note how

$$\frac{\partial(Y(\mathbf{x},\mathbf{z}))}{\partial \mathbf{z}} = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'.$$

[11] This section has somewhat more advanced material and may be skipped on a first reading.
[12] Sections 9.5 and 9.6 are based on Miro and del Castillo [104].

Thus, from (9.9) we can see that to minimize the variance of Y we need to make this partial derivative as small as possible, ideally making it equal to zero. This is directly related to our observation (Section 9.2) that the slope of the response with respect to the noise factors should be minimized to minimize the variance of Y.

In practice, all the parameters in model (9.1) are estimated from data. The model is linear in the parameters, and the ϵ_i's are i.i.d.. Therefore, all the usual assumptions of linear regression are met (see Appendix A).

To estimate $\beta_0,\ \boldsymbol{\beta},\ \mathbf{B},\ \boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$ using Ordinary Least Squares, it is useful to re-parameterize model (9.1) following a suggestion in Borror et al. [14]:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \tag{9.10}$$

where $\mathbf{y}$ is a vector containing all the observations $Y_i(\mathbf{x}_i, \mathbf{z}_i)$ and $\mathbf{X}^*$ is a matrix with $\frac{(k+1)(k+2)}{2} + kr + r$ columns representing: the linear effects, the second order interactions and quadratic effects for the controllable factors, the r linear effects for the noise variables, and the kr interactions between the noise and the controllable factors. Collecting all the terms associated with controllable factors and the intercept in one vector ($\boldsymbol{x}$ in "model form", hence the superscript):

$$\mathbf{x}^{(\mathbf{m})} = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_k & x_1x_2 & x_1x_3 & \dots & x_kx_{k-1} & x_1^2 & x_2^2 & \dots & x_k^2 \end{bmatrix} \tag{9.11}$$

The vector $\boldsymbol{\beta}^*$ contains all the parameters present in equation (9.1), $\beta_0,\ \boldsymbol{\beta},\ \mathbf{B},\ \boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$.

The least squares estimator of $\boldsymbol{\beta}^*$ is then given by:

$$\widehat{\boldsymbol{\beta}}^* = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{y}.$$

We will denote by $\widehat{Y}(\mathbf{x}, \mathbf{z})$ a prediction based on the fitted model (9.1) given specific values of controllable ($\mathbf{x}$) and noise ($\mathbf{z}$) factors.

The symmetric matrix $\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$, depicted in Figure 9.7, is the scaled covariance matrix of $\widehat{\boldsymbol{\beta}}^*$ and is made up of several submatrices relevant in our discussion. We also need to further partition $\mathbf{C}_{(\mathbf{zx})}$ into submatrices corresponding to each controllable factor $x_i,\ i = 1, \dots, k$. This partition is presented in Figure 9.8.

$$\left(\mathbf{X}^{*\prime}\mathbf{X}^{*}\right)^{-1} = \left[\begin{array}{c|c|c} \mathbf{C}_{\mathbf{x}(\mathbf{m})} & \mathbf{C}_{\mathbf{x}(\mathbf{m})\mathbf{z}} & \mathbf{C}_{\mathbf{x}(\mathbf{m})(\mathbf{zx})} \\ \hline \left(\mathbf{C}_{\mathbf{x}(\mathbf{m})\mathbf{z}}\right)' & \mathbf{C}_{\mathbf{z}} & \mathbf{C}_{\mathbf{z}(\mathbf{zx})} \\ \hline \left(\mathbf{C}_{\mathbf{x}(\mathbf{m})(\mathbf{zx})}\right)' & \left(\mathbf{C}_{\mathbf{z}(\mathbf{zx})}\right)' & \mathbf{C}_{(\mathbf{zx})} \end{array}\right]$$

$$\begin{array}{c|ccccccccccccccc} & 1 & x_1 & \cdots & x_k & x_1x_2 & \cdots & x_{k-1}x_k & x_1^2 & \cdots & x_k^2 & z_1 & \cdots & z_r & z_1x_1 & \cdots & z_1x_k & \cdots & z_rx_k \\ \hline 1 & & & & & & & & & & & & & & & & & & \\ x_1 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ x_k & & & & & & & & & & & & & & & & & & \\ x_1x_2 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ x_{k-1}x_k & & & & & & & & & & & & & & & & & & \\ x_1^2 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ x_k^2 & & & & & & & & & & & & & & & & & & \\ \hline z_1 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ z_r & & & & & & & & & & & & & & & & & & \\ \hline z_1x_1 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ z_rx_1 & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ z_1x_k & & & & & & & & & & & & & & & & & & \\ \vdots & & & & & & & & & & & & & & & & & & \\ z_rx_k & & & & & & & & & & & & & & & & & & \end{array}$$

Figure 9.7. The structure of the covariance matrix of $\widehat{\beta}^{*}$. From: [104]

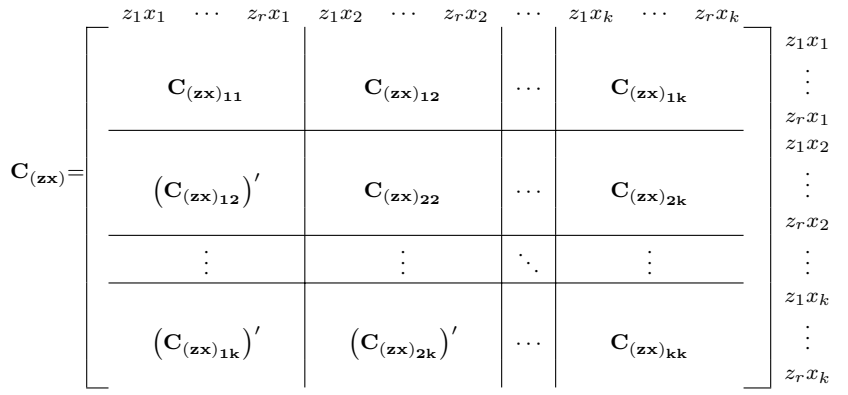


Figure 9.8. The structure of of $\mathbf{C}_{(\mathbf{zx})}$. From: [104]

To estimate $E_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ and $Var_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ some authors suggest to to simply substitute the parameters in equations (9.8) and (9.9) by their respective OLS estimators, $\widehat{\beta}_0$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{B}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Delta}}$, to obtain:

$$\widehat{E}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = \widehat{\beta_0} + \mathbf{x}'\widehat{\boldsymbol{\beta}} + \mathbf{x}'\widehat{\mathbf{B}}\mathbf{x} \tag{9.12}$$

$$\widehat{Var}_{\mathbf{z}}^{(b)}(Y(\mathbf{x},\mathbf{z})) = (\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}) + \widehat{\sigma}_{\epsilon}^2. \tag{9.13}$$

The estimator in equation (9.12) is an unbiased estimator of $E_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$. However, the estimator in equation (9.13) is not an unbiased estimator of $Var_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ because it is a quadratic function of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Delta}}$. Hence the "$(b)$" superscript. To show this, we can take expectation of the quadratic form in (9.13) over the parameter estimates $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Delta}}$ to get (see, for example, [3, p. 205]):

$$\begin{aligned}
&E_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}\left(\widehat{Var}_{\mathbf{z}}^{(b)}(Y(\mathbf{x},\mathbf{z}))\right) \\
&\quad = E_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}\left((\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}) + \widehat{\sigma}_{\epsilon}^2\right) \\
&\quad = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}) + tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}\, Var_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}\left(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}\right)\right) + \sigma_{\epsilon}^2 \\
&\quad = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}) + tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P}(\mathbf{x})\right) + \sigma_{\epsilon}^2
\end{aligned} \tag{9.14}$$

where the variance matrix of the slope $\mathbf{P}(\mathbf{x})$ is obtained by partitioning $\widehat{\boldsymbol{\Delta}}$ column-wise and assuming an orthogonal experimental design (more about this matrix below).

An unbiased estimator for $Var_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ can be found by correcting for the term with the trace in equation (9.14) in the following way:

$$\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = (\widehat{\gamma} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\gamma} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}) + \widehat{\sigma}_{\epsilon}^2\left(1 - trace(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P})\right). \tag{9.15}$$

This expression is the one frequently recommended as it is an unbiased estimator for $Var_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ (see [117]). Miro et al. [104] showed how an equivalent, but more revealing expression for $\mathbf{P}(\mathbf{x})$ can be found by partitioning $\widehat{\boldsymbol{\Delta}}$ row-wise instead of column-wise:

$$\begin{aligned}\mathbf{P}(\mathbf{x}) &= Var_{\widehat{\gamma},\widehat{\boldsymbol{\Delta}}}\left(\widehat{\gamma} + \left[\widehat{\boldsymbol{\delta}}'_{1\cdot}\widehat{\boldsymbol{\delta}}'_{2\cdot}\cdots\widehat{\boldsymbol{\delta}}'_{k\cdot}\right]\mathbf{x}\right)\Big/\sigma_{\epsilon}^2 \\ &= Var\left(\widehat{\gamma} + \sum_{i=1}^{k} x_i\widehat{\boldsymbol{\delta}}'_{i\cdot}\right)\Big/\sigma_{\epsilon}^2.\end{aligned}$$

where each $\widehat{\boldsymbol{\delta}}'_{i\cdot}$ is a $r\times 1$ vector that contains the columns of $\boldsymbol{\Delta}$. The variance of the slope has simpler forms under some assumptions, which we now consider.

If the DOE renders i) N and $N\times C$ effects orthogonal, and ii) $N\times C$ effects orthogonal to other $N\times C$ effects, we get:

$$\mathbf{P}(\mathbf{x}) = \mathbf{C}_{\mathbf{z}} + \sum_{i=1}^{k} x_i^2\mathbf{C}_{(\mathbf{zx})_{ii}}$$

where the scaled covariance matrices of $\widehat{\gamma}$ and $\widehat{\boldsymbol{\delta}}_{i\cdot}$ are given by $\mathbf{C}_{\mathbf{z}}$ (Figure 9.7) and $\mathbf{C}_{(\mathbf{zx})_{ii}}$ (Figure 9.8), respectively. Here, a scaled covariance matrix means a covariance matrix that is divided by σ_{ϵ}^2.

An even simpler form of the variance of the slope can be obtained as follows. If the DOE renders all the $N\times C$ interactions with the same standard error[13] then all the $\widehat{\boldsymbol{\delta}}_{i\cdot}$'s have the same covariance matrix. Therefore, let $\mathbf{C}_{\delta} = \mathbf{C}_{(\mathbf{zx})_{ii}}\ \forall i = 1\ldots k$, denote this common covariance matrix. Then we have that $\mathbf{P}(\mathbf{x})$ reduces even further to:

$$\mathbf{P}(\mathbf{x}) = \mathbf{C}_{\mathbf{z}} + (\mathbf{x}'\mathbf{x})\mathbf{C}_{\delta}.$$

[13]Most of the response surface experimental designs, especially of the mixed resolution type that have been recommended for use in the dual response approach to RPD [15], [13], [117] have this property, see Section 9.3.

In this case, an alternative expression of the unbiased estimator of the variance can be found as:

$$\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = (\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}) + \widehat{\sigma}^2_{\epsilon}\left(1 - \left(\underbrace{tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{C}_{\mathbf{z}}\right) + \left(\mathbf{x}'\mathbf{x}\right) tr(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{C}_{\delta})}_{\text{Bias correction term}}\right)\right). \quad (9.16)$$

The unbiased estimator in (9.16) explicitly states how the "bias correction term" increases with the squared distance from the origin in coded units ($\mathbf{x}'\mathbf{x}$). Thus, for settings of the controllable factors far away from the origin, the variance estimate (9.16) can be negative. This can also happen if the part of the model containing noise factors does not fit well. We now illustrate some of the aforementioned computations which will be used in later sections in this chapter.

Example. Computation of $\mathbf{P}(\mathbf{x})$ and $(\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}$. Borror and Montgomery [15] analyze the data set presented in Table 9.8, which corresponds to a CMR design with $k = 4$ and $r = 3$. The fitted response surface has an R^2 statistic of 0.945, which implies a relatively good fit. We illustrate the computation of the $(\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}$ and the $\mathbf{P}(\mathbf{x})$ matrix with this design.

Table 9.9 shows the corresponding $(\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}$ matrix, indicating which terms each column and row corresponds to. Let us assume that from historical experience, the noise factors have a variance covariance matrix equal to:

$$\boldsymbol{\Sigma}_{\mathbf{z}} = \frac{1}{7.5}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since all $C \times N$ interactions have the same standard error, then (see Table 9.9) we have that $\mathbf{C}_{\delta} = 0.03125\mathbf{I}_3 = \mathbf{C}_{(\mathbf{zx}_{ii})}$ for $i = 1, 2, 3$. Therefore, the covariance matrix of slopes is given by

$$\mathbf{P} = \mathbf{C}_{\mathbf{z}} + (\mathbf{x}'\mathbf{x})\mathbf{C}_{\delta} = \begin{bmatrix} \frac{1}{7.5} + 0.01325\sum_{i=1}^{4} x_i^2 & 0 & 0 \\ 0 & \frac{1}{7.5} + 0.01325\sum_{i=1}^{4} x_i^2 & 0 \\ 0 & 0 & \frac{1}{7.5} + 0.01325\sum_{i=1}^{4} x_i^2 \end{bmatrix}.$$

These computations will be used in the example shown in the next section. ■

Table 9.8. Composite Mixed Resolution Design for the example (from: Borror et al. [14])

x_1	x_2	x_3	x_4	z_1	z_2	z_3	Y
1	1	1	1	1	1	1	411
−1	1	1	1	1	−1	−1	269
1	−1	1	1	1	−1	−1	106
−1	−1	1	1	1	1	1	159
1	1	−1	1	1	−1	−1	258
−1	1	−1	1	1	1	1	257
1	−1	−1	1	1	1	1	216
−1	−1	−1	1	1	−1	−1	97
1	1	1	−1	1	1	−1	312
−1	1	1	−1	1	−1	1	150
1	−1	1	−1	1	−1	1	253
−1	−1	1	−1	1	1	−1	111
1	1	−1	−1	1	−1	1	56
−1	1	−1	−1	1	1	−1	35
1	−1	−1	−1	1	1	−1	204
−1	−1	−1	−1	1	−1	1	−47
1	1	1	1	−1	−1	1	175
−1	1	1	1	−1	1	−1	169
1	−1	1	1	−1	1	−1	152
−1	−1	1	1	−1	−1	1	163
1	1	−1	1	−1	1	−1	143
−1	1	−1	1	−1	−1	1	222
1	−1	−1	1	−1	−1	1	180
−1	−1	−1	1	−1	1	−1	136
1	1	1	−1	−1	−1	−1	109
−1	1	1	−1	−1	1	1	188
1	−1	1	−1	−1	1	1	174
−1	−1	1	−1	−1	−1	−1	164
1	1	−1	−1	−1	1	1	85

Table 9.8. (Continued)

−1	1	−1	−1	−1	−1	−1	159
1	−1	−1	−1	−1	−1	−1	154
−1	−1	−1	−1	−1	1	1	67
2.65	0	0	0	0	0	0	17
−2.65	0	0	0	0	0	0	−54
0	2.65	0	0	0	0	0	398
0	−2.65	0	0	0	0	0	194
0	0	2.65	0	0	0	0	370
0	0	−2.65	0	0	0	0	237
0	0	0	2.65	0	0	0	379
0	0	0	−2.65	0	0	0	255
0	0	0	0	0	0	0	89
0	0	0	0	0	0	0	9
0	0	0	0	0	0	0	69

9.6 Robustness with respect to Noise Factors and Parameter Uncertainty

If $\widehat{Var}_z^{(b)}(Y(\mathbf{x},\mathbf{z}))$ is minimized in a Robust Parameter Design problem, we will be focusing only on the variance due to the noise factors. However, an additional variance component not only in RPD but in any response surface problem is due to the *uncertainty in the parameter estimates*. This was discussed in Chapters 6 and 7 without reference to RPD problems. In the context of RPD, imprecise estimates of the parameters associated with noise factors can result in negative values of $\widehat{Var}_z^{(b)}(Y(\mathbf{x},\mathbf{z}))$. Therefore, a natural extension of the dual response approach for RPD shown in previous sections is to introduce the additional variance of the parameters estimates into an objective function that combines it with noise factor variance. Optimizing such function will achieve a process that is robust with respect to noise factor variation *and* with respect to uncertainty in the parameter estimates.

One such function is the *variance of the predicted response*, where the variance is taken with respect to the parameter estimates of the model *and* with respect to the noise factors. As in the previous section, let $\widehat{\boldsymbol{\beta}}^*$ denote the vector containing all the parameters in the model of equation (9.1).

Table 9.9. The $(\mathbf{X}^{*'}\mathbf{X}^{*})^{-1}$ matrix for the data set in Table 9.8

1	x_1	...	x_4	x_1x_2	...	x_3x_4	x_1^2	x_2^2	x_3^2	x_4^2	z_1	z_2	z_3	z_1x_1	z_2x_1	...	z_2x_4	z_3x_4	
0.18	0		0	0		0	−0.04	−0.04	−0.04	−0.04	0	0	0	0	0		0	0	1
0	0.02		0	0		0	0	0	0	0	0	0	0	0	0		0	0	x_1
		⋱			⋱											⋱			⋮
0	0		0.02	0		0	0	0	0	0	0	0	0	0	0		0	0	x_4
0	0		0	0.03		0	0	0	0	0	0	0	0	0	0		0	0	x_1x_2
		⋱			⋱											⋱			⋮
0	0		0	0		0.03	0	0	0	0	0	0	0	0	0		0	0	x_3x_4
−0.04	0		0	0		0	0.02	0.01	0.01	0.01	0	0	0	0	0		0	0	x_1^2
−0.04	0		0	0		0	0.01	0.02	0.01	0.01	0	0	0	0	0		0	0	x_2^2
−0.04	0		0	0		0	0.01	0.01	0.02	0.01	0	0	0	0	0		0	0	x_3^2
−0.04	0		0	0		0	0.01	0.01	0.01	0.02	0	0	0	0	0		0	0	x_4^2
0	0		0	0		0	0	0	0	0	0.03	0	0	0	0		0	0	z_1
0	0		0	0		0	0	0	0	0	0	0.03	0	0	0		0	0	z_2
0	0		0	0		0	0	0	0	0	0	0	0.03	0	0		0	0	z_3
0	0		0	0		0	0	0	0	0	0	0	0	0.03	0		0	0	z_1x_1
0	0		0	0		0	0	0	0	0	0	0	0	0	0.03		0	0	z_2x_1
		⋱			⋱											⋱			⋮
0	0		0	0		0	0	0	0	0	0	0	0	0	0		0.03	0	z_2x_4
0	0		0	0		0	0	0	0	0	0	0	0	0	0		0	0.03	z_3x_4

Miro and del Castillo [104] propose to consider the objective:

$$Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}(\widehat{Y}(\mathbf{x},\mathbf{z})) = Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}\left(\widehat{\beta_0} + \mathbf{x}'\widehat{\boldsymbol{\beta}} + \mathbf{x}'\widehat{\mathbf{B}}\mathbf{x} + \mathbf{z}'\widehat{\boldsymbol{\gamma}} + \mathbf{x}'\widehat{\boldsymbol{\Delta}}\mathbf{z}\right) \quad (9.17)$$

which is the variance of the predicted response taking into consideration the variability in the noise factors and in the parameter estimates.

If the parameter estimates associated with the controllable and noise factors are orthogonal, then

$$Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}(\widehat{Y}(\mathbf{x},\mathbf{z})) = Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}\left(\widehat{\beta_0} + \mathbf{x}'\widehat{\boldsymbol{\beta}} + \mathbf{x}'\widehat{\mathbf{B}}\mathbf{x}\right) + Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}\left(\mathbf{z}'\widehat{\boldsymbol{\gamma}} + \mathbf{x}'\widehat{\boldsymbol{\Delta}}\mathbf{z}\right) \quad (9.18)$$

Let $\mathbf{x}^{(\mathbf{m})}$ denote the $\frac{(k+1)(k+2)}{2}$ vector containing the regressors associated with the controllable factors. Then, since the argument of the variance operator in the first term of (9.18) is not a function of $\mathbf{z}$, this term is just the variance with respect to $\widehat{\boldsymbol{\beta}}^{*}$, given by:

$$Var_{\widehat{\boldsymbol{\beta}}^{*}}\left(\widehat{\beta_0} + \mathbf{x}'\widehat{\boldsymbol{\beta}} + \mathbf{x}'\widehat{\mathbf{B}}\mathbf{x}\right) = \sigma_{\epsilon}^{2}\mathbf{x}'^{(\mathbf{m})}\mathbf{C}_{\mathbf{x}^{(\mathbf{m})}}\mathbf{x}^{(\mathbf{m})} \quad (9.19)$$

where $\mathbf{C}_{\mathbf{x}^{(\mathbf{m})}}$ is the upper left matrix of $\left(\mathbf{X}^{*'}\mathbf{X}^{*}\right)^{-1}$, shown in Figure 9.7.

The second term of equation (9.18) can be found to be equal to (see Problem 2):

$$Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}\left(\mathbf{z}'\widehat{\boldsymbol{\gamma}} + \mathbf{x}'\widehat{\boldsymbol{\Delta}}\mathbf{z}\right) = \sigma_{\epsilon}^{2}tr(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P}) + (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}). \quad (9.20)$$

where tr denotes the trace (see Appendix C). Therefore, combining equations (9.19) and (9.20) gives the variance of the predicted response:

$$\begin{aligned} &Var_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}(\widehat{Y}(\mathbf{x},\mathbf{z})) \\ &\quad = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}) + \sigma_{\epsilon}^{2}\left(tr(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P}) + \mathbf{x}'^{(\mathbf{m})}\mathbf{C}_{\mathbf{x}^{(\mathbf{m})}}\mathbf{x}^{(\mathbf{m})}\right). \end{aligned} \quad (9.21)$$

We need an estimator of this. An unbiased estimator was found by Miro et al. as follows. From equation (9.14) we have that:

$$E_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}\left((\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\boldsymbol{\gamma}} + \widehat{\Delta}'\mathbf{x})\right) = (\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x}) + \sigma_{\epsilon}^{2}tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P}\right) \quad (9.22)$$

and therefore an unbiased estimator of (9.20) is given by:

$$\widehat{Var}_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}(\widehat{Y}(\mathbf{x},\mathbf{z})) = (\widehat{\boldsymbol{\gamma}}+\widehat{\boldsymbol{\Delta}}'\mathbf{x})'\boldsymbol{\Sigma}_{\mathbf{z}}(\widehat{\boldsymbol{\gamma}}+\widehat{\boldsymbol{\Delta}}'\mathbf{x})+\widehat{\sigma}^{2}_{\epsilon}(\mathbf{x}'^{(\mathbf{m})}\mathbf{C}_{\mathbf{x}^{(\mathbf{m})}}\mathbf{x}^{(\mathbf{m})}) \quad (9.23)$$

(see Problem 3). This is the function proposed in [104] to solve RPD problems. That is, *we minimize the estimated variance of the predictions*. As indicated by the subscripts, (9.23) considers the variability created both by the randomness of the noise factors and by the uncertainty in the parameter estimates of the fitted model.

Example. RPD based on the estimated variance of the predictions. Let us consider again the data in Table 9.8. Assume as before that covariance matrix of the noise factors is $\boldsymbol{\Sigma}_{\mathbf{z}} = \frac{1}{7.5}\mathbf{I}_3$. The goal is to obtain process settings such that the target of 200 units is achieved with minimum variance. In this example we will compare the estimated prediction variance, $\widehat{Var}_{\mathbf{z},\,\widehat{\boldsymbol{\beta}}^{*}}(\widehat{Y}(\mathbf{x},\mathbf{z}))$, computed using equation (9.23), with the biased estimator $\widehat{Var}^{(b)}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ from equation (9.13) and with $\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ from equation (9.16). These objectives were minimized subject to $\widehat{E}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = 200$ and $\mathbf{x}'\mathbf{x} \leq 7$ (this is the square of the radius of the experimental region). Matlab's fmincon routine was used for this purpose, although given the small number of decision variables any general purpose non-linear programming technique can be utilized, such as the Generalized Reduced Gradient (GRG) included in MS Excel's Solver.

The optimal solutions for the three constrained minimization problems are presented in Table 9.10. The column under $\mathbf{x}^{*}$ contains the optimal settings of the controllable factors for the respective objective. The value of the objectives evaluated at their respective $\mathbf{x}^{*}$ are listed under "Optimal Value". The values of the other objective functions for each optimal solution are also included in the following three columns. $\widehat{\sigma}_{\widehat{E}_{\mathbf{z}}(Y)}$ denotes the estimated standard error with which the estimated expected mean of $\widehat{E}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z})) = 200$ is predicted. These values correspond to $\widehat{\sigma}_{\epsilon}\sqrt{\mathbf{x}'^{(\mathbf{m})*}\mathbf{C}_{\mathbf{x}^{(\mathbf{m})}}\mathbf{x}^{(\mathbf{m})*}}$, where $\mathbf{x}'^{(\mathbf{m})*}$ denotes the optimal settings $\mathbf{x}^{*}$ transformed to model form.

As it can be seen, the criterion $\widehat{Var}(\widehat{Y}(\mathbf{x},\mathbf{z}))$ provides the smallest standard error around the desired mean. This implies that the third solution on the table is the most robust with respect to i) noise factor variation, and ii) parameter estimation variability. Notice how from the R^2 value (0.945) one would think that ii) above would have little impact, when in fact *the effect of the parameter estimation variability is very dramatic*. ■

Table 9.10. Solutions obtained from minimizing $\widehat{Var}_{\mathbf{z}}^{(b)}(Y(\mathbf{x},\mathbf{z}))$, $\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$ and $\widehat{Var}_{\mathbf{z},\widehat{\beta}^*}(\widehat{Y}(\mathbf{x},\mathbf{z}))$, RPD example. When minimizing $\widehat{Var}^{(b)}(\widehat{Y}(\mathbf{x},\mathbf{z}))$, multiple solutions were obtained. Source: [104]

Objective Function	$\mathbf{x}^*$	$\widehat{Var}_{\mathbf{z}}^{(b)}(Y(\mathbf{x}^*,\mathbf{z}))$	$\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x}^*,\mathbf{z}))$	$\widehat{Var}_{\mathbf{z},\widehat{\beta}^*}(\widehat{Y}(\mathbf{x}^*,\mathbf{z}))$	$\widehat{\sigma}_{\widehat{E}_{\mathbf{z}}(Y)}$
$\widehat{Var}_{\mathbf{z}}^{(b)}(Y(\mathbf{x},\mathbf{z}))$	$x_1 = -0.73 + 1.01x_4$ $x_2 = \ \ 0.59 - 0.47x_4$ $x_3 = -0.22 - 1.96x_4$ $x_4 \in [-0.90\,,\ 1.10]$	1868	1710	1018	depends on $\mathbf{x}^*$
$\widehat{Var}_{\mathbf{z}}(Y(\mathbf{x},\mathbf{z}))$	$[-1.86,\ 1.18,\ 1.29, -0.70]$	1874	1687	1478	38.4
$\widehat{Var}_{\mathbf{z},\widehat{\beta}^*}(\widehat{Y}(\mathbf{x},\mathbf{z}))$	$[\,0.18,\ 0.46, -1.28,\ 0.96]$	1890	1801	370	18.6

Compared with the solutions found with the other variance objectives, the new objective allows us to find solutions for which the variance contribution due to the noise factors is *also* small but *in addition* the predictive properties of the fitted models are much better. Thus, we obtain a solution that is robust to both noise factor variability and uncertainty in the model parameter estimates.

9.7 The Noise Factor Separation Criterion for Experimental Designs Used in Robust Parameter Design

In a recent paper [40] it was proposed to evaluate experimental designs used for RPD based on how much one needs to separate the two levels of the noise factors. The idea is that if noise factor levels are separated little, little or no information on the model parameters associated with them will be available. The problem, in practice, is that wider noise factor levels, while convenient from a model-fitting point of view, are frequently more expensive than narrower ones. For example, a typical noise factor in machining and other discrete-part manufacturing processes is environmental temperature. Varying environmental temperature in a RPD experiment over wide ranges will be more expensive than using narrower ranges, given the increase in energy expense. As a second example, consider experiments in a wind tunnel to design car components. Wind speed is a noise factor that can be varied in such facility, with wider ranges always associated with increasing cost. Even in case an increasing range in the noise factor does not result in a significant increment in cost, there may be a maximum range beyond which no experiment can be run (environmental temperature and wind speed are also instances of such case). In addition, setting the factor levels too far apart may result in model bias, i.e., over the experimental region a more nonlinear model than assumed may hold (this is true for all factors, not only for the noise factors). In summary, there is a trade-off between the goodness of the fit of the model (in particular, the part that depends on the noise factors $\mathbf{z}$), and the size of the experimental region in the noise factor space. The noise factor separation (**NFS**) balances these two aspects of the trade-off.

Suppose we wish to fit model (9.1). The terms in the model that contain noise factors $(\mathbf{z})$ will be referred to as the *noise part of the model*. Usual practice in response surface approaches to RPD is to let $z_i = \pm 1$ represent plus or minus one (or perhaps, two) standard deviation(s) of the distribution of the

noise factors in original units (see [117, p. 562]). The NFS design criterion is based on preferring designs that, other things being equal, require less extreme noise factor levels in original units, i.e., less scaling.

We now explain how to compute the noise factor separation $C_D(\tau)$ that achieves a given expected mean square error (MSE) τ of the noise part of the model for a given experimental design D. Following Section 9.5, let

$$\mathbf{P}(\mathbf{x}) = Var_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}\left(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x}\right). \tag{9.24}$$

This is the $r \times r$ covariance matrix of the estimated slope of model (9.1) in the direction of the r noise factors (i.e., Var$(\partial Y/\partial \mathbf{z})$), taken with respect to the estimated parameters [14]. This quantity determines how well we estimate the variance response (9.9), something essential for the success of any RPD study [117]. It is also related to the quantities plotted on Variance Dispersion Graphs for the slope [14], which plot the max, min, and average of the diagonal elements of $\mathbf{P}(\mathbf{x})$ (i.e., $\mathbf{P}(\mathbf{x})_{ii}$ for $i = 1, 2, \ldots, r$) over spheres of different radii in the controllable factor space.

The covariance matrix of the slopes relates to the expected mean square error (MSE) of the noise part of the model as follows:

$$\begin{aligned} \text{MSE}((\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\mathbf{z}/\sigma_\epsilon^2) &= Var_{\widehat{\boldsymbol{\gamma}},\widehat{\boldsymbol{\Delta}}}((\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\mathbf{z})/\sigma_\epsilon^2 \\ &= \mathbf{z}'\mathbf{P}(\mathbf{x})\mathbf{z} \end{aligned}$$

where the first equality follows from the fact that $(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\mathbf{z}$ is an unbiased estimator of $(\boldsymbol{\gamma} + \boldsymbol{\Delta}'\mathbf{x})'\mathbf{z}$ and hence its variance equals the MSE.

As shown by Miro and del Castillo [104], the MSE of $(\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\mathbf{z}/\sigma_\epsilon^2$, averaged over all possible values of the noise factors $\mathbf{z}$ is then given by:

$$\begin{aligned} E_{\mathbf{z}}\left(\text{MSE}((\widehat{\boldsymbol{\gamma}} + \widehat{\boldsymbol{\Delta}}'\mathbf{x})'\mathbf{z}/\sigma_\epsilon^2)\right) &= E_{\mathbf{z}}(\mathbf{z}'\mathbf{P}(\mathbf{x})\mathbf{z}) \\ &= tr(\boldsymbol{\Sigma}_{\mathbf{z}}\mathbf{P}(\mathbf{x})) \end{aligned} \tag{9.25}$$

assuming, without loss of generality, that the coded noise factors have zero mean. This is the case if we introduce the following coding convention of the noise factors:

$$z_i = \frac{\xi_i - \mu_{\xi_i}}{C\sigma_{\xi_i}}, \quad \forall i = 1, \ldots, r$$

where ξ_i is the uncoded noise factor which has mean μ_i and variance $\sigma^2_{\xi_i}$ and C is a scaling factor. Then,

$$E(z_i) = \frac{\mu_{\xi_i} - \mu_{\xi_i}}{C\sigma_{\xi_i}} = 0, \qquad Var(z_i) = \frac{\sigma^2_{\xi_i}}{C^2\sigma^2_{\xi_i}} = \frac{1}{C^2}, \qquad \forall i = 1, \ldots, r \ .$$

or, in vector notation:

$$E(\mathbf{z}) = 0 \ , \qquad Var(\mathbf{z}) = \mathbf{\Sigma_z} = \frac{1}{C^2}\mathbf{\Psi}_\xi \tag{9.26}$$

where $\mathbf{\Psi}_\xi$ is the correlation matrix of the noise variables measured in their original units.

With this coding convention, a coded factor of $z_i = \pm 1$ corresponds to original (uncoded) noise factors *separated* $\pm C$ standard deviations apart. Thus, larger values of C indicate more scaling and more noise factor separation in uncoded units. From (9.25) and (9.26) the coding convention implies that the expected MSE of the noise part of the model is:

$$E_{\mathbf{z}}\left(\text{MSE}((\widehat{\gamma} + \widehat{\mathbf{\Delta}}'\mathbf{x})'\mathbf{z}/\sigma^2_\epsilon)\right) = \frac{1}{C^2} tr(\mathbf{\Psi}_\xi \mathbf{P}(\mathbf{x})). \tag{9.27}$$

In general, the expected MSE increases the farther the point $\mathbf{x}$ at which we want to predict is from the origin. Therefore, it was suggested in [104] to look at the expected MSE at $\mathbf{x}_{max}$, the design point farthest away from the origin of the controllable factor space. If τ represents the desired expected MSE for the noise part of the model evaluated at $\mathbf{x}_{max}$, the value of C that achieves it, $C_D(\tau)$, is obtained from(9.27) equating the right hand side to τ and solving for C:

$$C_D(\tau) = \sqrt{\frac{1}{\tau} tr(\mathbf{\Psi}_\xi \mathbf{P}(\mathbf{x_{max}}))}. \tag{9.28}$$

Values of the scaling parameter C larger than the right hand side of (9.28) will give expected MSE values lower than τ. The relation given by (9.28) is a hyperbola that always has the shape indicated in Figure 9.9. Since the quantity inside the trace in (9.28) is not a function of τ, if a design has smaller $C_D(\tau)$ for $\tau = \tau_0$, it will have smaller C value for all τ. Therefore, one can simply compare designs according to the NFS criterion by looking at their performance at a single value of the expected MSE of the noise part of the model, τ, on Figure 9.9.

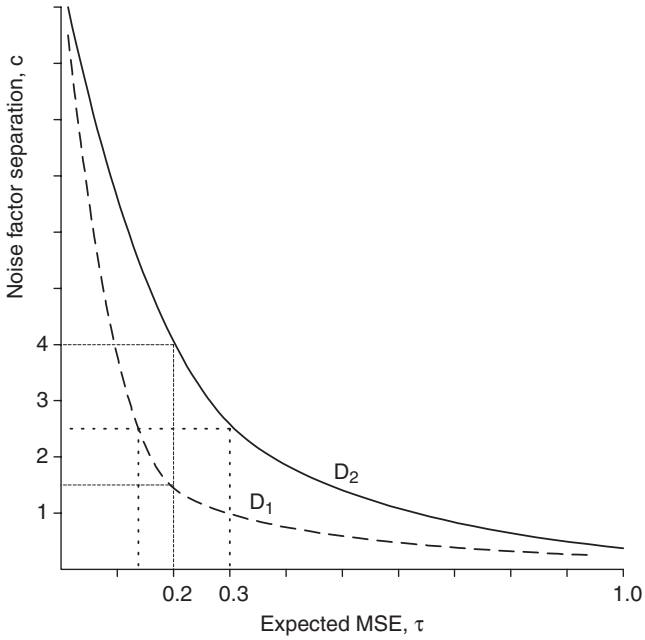


Figure 9.9. Illustration of the relation between the expected MSE of the noise part of the model and the noise factor separation (scaling) factor C for two hypothetical designs. Source: [40]

To illustrate how an experimental design can be evaluated using the NFS criterion $C_D(\tau)$, suppose we have two designs, D_1 and D_2. One can vary the amount of expected MSE desired by the experimenter for the noise part of the model, τ, and see what noise factor separation C achieves such expected MSE. Figure 9.9 illustrates this idea for two hypothetical designs. The shape of the functions, which is always as in the figure, indicates that as the expected MSE we desire goes down, the necessary scaling factor goes up, and with it, the separation between the noise factors. For example, let us suppose we wish to have an expected MSE of the noise part of the model of at most 20% the uncontrollable noise variance σ_ϵ^2, thus $\tau = 0.2$. To achieve this expected MSE, design D_1 requires a noise factor separation of $C = 1.5$ standard deviations, while design D_2 provides the same expected MSE of $0.2\sigma^2$ for a noise factor separation of $C = 4$ standard deviations (thinner dotted lines). Thus, we should prefer design D_1 to design D_2 (or, we shall say design D_1 is better than design D_2 according to the *noise factor separation (NFS) criterion*).

An alternative and also useful interpretation of Figure 9.9 is the following. Suppose we have set the noise factors 2.5 standard deviations apart. If we conduct experiment D_1, our expected MSE for the noise part of the model is around $0.14\sigma_\epsilon^2$. In contrast, if we conduct the experiment according to experimental design D_2 the MSE we should expect for the noise part of the model is $0.3\sigma_\epsilon^2$ (thicker dotted line). Thus, D_2 is inferior to D_1 according to the NFS criterion.

It can be shown easily (see Problem 5) that the $C(\tau)$ functions of two different designs will never "cross" (they can only coincide for the same DOE). Therefore, it is not necessary to draw graphs like Figure 9.9 to compare designs; a design D_1 that is better (i.e., gives lower noise factor separation) than other design for a given value of the expected MSE will be better for *all* expected values of the MSE. Therefore, it suffices to compare the designs at a single τ-value of the graph. del Castillo et al. [40] used the value $C_D(0.1)$, that is, the noise factor separation required to get an expected MSE of the noise part of the model of 0.1 times σ_ϵ^2.

The noise factor separation criteria $C_D(\tau)$ depends on several aspects we can change in a design D:

- the farthest controllable factor point in the design $\mathbf{x}_{max}$;
- how well D estimates the $N \times C$ interaction effects;
- the correlation between $N \times C$ and N effects and between each $N \times C$ and other $N \times C$ effects;
- the design size (n).

del Castillo et al. [40] propose D-optimal designs for robust parameter experiments subject to a NFS constraint. They also investigated some well-known designs from a NFS point of view. A summary of their findings is as follows:

- Box-Behnken (BB) designs. These designs should not be used to fit model (9.1) given the poor standard error with which they estimate the model

interactions. This results in larger NFS values and low D and G efficiencies than other much smaller designs.

- CMR and modified CCD designs. With few exceptions, CMR designs require smaller noise factor separation than other designs to provide an expected MSE for the noise part of the model of $0.1\sigma_\epsilon^2$ at the farthest point $\mathbf{x}$. They almost always provide highest D-efficiency among the designs on the list. These designs coincide with modified CCD's for cases $(k, r) = (2, 2), (3, 2), (3, 3), (5, 3), (7, 4)$, and $(8, 3)$.
- Smaller designs: Small Composite Designs (SCD's) and Hybrid (Roque more). It was concluded that the SCD's should not be used to fit model (9.1). They require very large noise factor separation to improve the precision of the estimation of the noise part of the model. The same can be said of the hybrid designs. These designs require considerable larger noise factor separation to give good precision in the noise part of the model.

For some new designs for RPD obtained using the NFS criterion in conjunction with other criteria, see [40].

9.8 Multiple Response Robust Parameter Design**

As with most real optimization problems[14], robust parameter design problems have multiple responses of interest. Here we present an approach due to Miro and del Castillo [105] who extended the Dual Response approach to RPD to the case of multiple responses.

Define a regression model with q responses, k controllable factors and r noise variables in the following way:

$$\underbrace{\mathbf{Y}}_{q\times 1} = \underbrace{\mathbf{\Theta}'}_{q\times p}\underbrace{\mathbf{x}^{(\mathbf{m})}}_{p\times 1} + \underbrace{\mathbf{\Delta}'}_{q\times r(k+1)}\underbrace{\mathbf{z}^{(\mathbf{m})}}_{r(k+1)\times 1} + \underbrace{\boldsymbol{\epsilon}}_{q\times 1} \tag{9.29}$$

where $\mathbf{\Theta}$ is a $p \times q$ matrix of coefficients for the controllable factors (each column contains all the coefficients for one response) and $\mathbf{x}^{(\mathbf{m})}$ is a $p \times 1$ vector containing the regressors for the controllable factors in model form (there are $p = \frac{(k+1)(k+2)}{2}$ such factors for a full quadratic model; as before, $(\mathbf{m})$ stands for $\mathbf{x}$ in "model form"):

[14]This section, based on reference [105], contains somewhat more advanced material and may be skipped on a first reading.

$$\mathbf{x^{(m)}} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_k \\ x_1x_2 \\ x_1x_3 \\ \vdots \\ x_{k-1}x_k \\ x_1^2 \\ x_2^2 \\ \vdots \\ x_k^2 \end{bmatrix}. \tag{9.30}$$

The matrix $\mathbf{\Delta}$ is $r(k+1) \times q$ and contains the first order coefficients for the noise factors as well as coefficients for the controllable factor-noise interactions. These terms are combined in a single matrix to avoid obtaining a cross-covariance term when applying the variance operator. The matrix is:

$$\mathbf{\Delta} = \begin{bmatrix} \beta_{y_1z_1} & \beta_{y_2z_1} & \cdots & \beta_{y_qz_1} \\ \beta_{y_1z_1x_1} & \beta_{y_2z_1x_1} & \cdots & \beta_{y_qz_1x_1} \\ \beta_{y_1z_1x_2} & \beta_{y_2z_1x_2} & \cdots & \beta_{y_qz_1x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{y_1z_1x_k} & \beta_{y_2z_2x_k} & \cdots & \beta_{y_qz_2x_k} \\ \beta_{y_1z_2} & \beta_{y_2z_2} & \cdots & \beta_{y_qz_2} \\ \beta_{y_1z_2x_1} & \beta_{y_2z_2x_1} & \cdots & \beta_{y_qz_2x_1} \\ \beta_{y_1z_2x_2} & \beta_{y_2z_2x_2} & \cdots & \beta_{y_qz_2x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{y_1z_2x_k} & \beta_{y_2z_2x_k} & \cdots & \beta_{y_qz_2x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{y_1z_r} & \beta_{y_2z_r} & \cdots & \beta_{y_qz_r} \\ \beta_{y_1z_rx_1} & \beta_{y_2z_rx_1} & \cdots & \beta_{y_qz_rx_1} \\ \beta_{y_1z_rx_2} & \beta_{y_2z_rx_2} & \cdots & \beta_{y_qz_rx_2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{y_1z_rx_k} & \beta_{y_2z_rx_k} & \cdots & \beta_{y_qz_rx_k} \end{bmatrix} \tag{9.31}$$

where $\beta_{y_ix_jz_k}$ is the coefficient for the interaction between x_j and z_k for the i^{th} response. The first order coefficients for the noise factors are the $\beta_{y_iz_k}$, hence

the matrix is $r(k+1)\times q$ and compatible with the definition of $\mathbf{z}^{(\mathbf{m})}$ that follows next.

The vector $\mathbf{z}^{(\mathbf{m})}$ is $r(k+1)\times 1$ and contains the noise factors *and their interactions with the controllable factors*:

$$\mathbf{z}^{(\mathbf{m})} = \begin{bmatrix} z_1 \\ z_1x_1 \\ z_1x_2 \\ \vdots \\ z_1x_k \\ z_2 \\ z_2x_1 \\ z_2x_2 \\ \vdots \\ z_2x_k \\ \vdots \\ z_r \\ z_rx_1 \\ z_rx_2 \\ \vdots \\ z_rx_k \end{bmatrix}. \tag{9.32}$$

Finally, the vector ϵ is $q\times 1$ and assumed normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_\epsilon$, that is,

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_q \end{bmatrix} \sim N_q(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$$

Notice that equation (9.29) assumes that all the responses can be appropriately modelled by functions of the same form, that is, by functions containing the same significant parameters.

Miro et al. assume that

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_r \end{bmatrix} \sim N_r(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{z}})$$

with $\boldsymbol{\Sigma}_{\mathbf{z}}$ known, possibly from historical data, similarly as in the univariate RPD problem.

Using the previous assumption one can take expectation and variance operators in equation (9.29):

$$E(\mathbf{Y}) = \boldsymbol{\Theta}' \mathbf{x}^{(\mathbf{m})}. \tag{9.33}$$

A multivariate approach equivalent to the one used by Box-Jones [27] and Myers [113] for the univariate case is to take the variance operator in equation (9.29) to obtain an equation useful to predict the variance of the responses once the noise factors are not tightly controlled as they were during the experiments. In other words we would get the $q \times q$ matrix:

$$Var(\mathbf{Y}) = \boldsymbol{\Delta}' Cov(\mathbf{z}^{(\mathbf{m})}) \boldsymbol{\Delta} + \boldsymbol{\Sigma}_{\epsilon} \tag{9.34}$$

where $Cov(\mathbf{z}^{(\mathbf{m})})$ is a $[(k+1)r \times (k+1)r]$ matrix shown in Figure 9.10.

The terms $\sigma_{z_i z_j}$ denote the covariance between z_i and z_j for $i \neq j$ and $\sigma^2_{z_i}$ denotes the variance of z_i. Note how if the noise factors are uncorrelated, $Cov(\mathbf{z}^{(\mathbf{m})})$ is block diagonal. The matrix $Cov(\mathbf{z}^{(\mathbf{m})})$ can be written as:

$$Cov(\mathbf{z}^{(\mathbf{m})}) = \boldsymbol{\Sigma}_{\mathbf{z}} \otimes \left[\mathbf{x}^{(\mathbf{l})} \mathbf{x}'^{(\mathbf{l})} \right] \tag{9.35}$$

where $\mathbf{x}^{(\mathbf{l})}$ is formed by the first $k+1$ elements of $\mathbf{x}^{(\mathbf{m})}$, $\boldsymbol{\Sigma}_{\mathbf{z}}$ is the covariance matrix of the noise factors and $\otimes$ denotes the Kronecker or direct product. Therefore, we have that:

$$Var(\mathbf{Y}) = \boldsymbol{\Delta}' \left[\boldsymbol{\Sigma}_{\mathbf{z}} \otimes \left(\mathbf{x}^{(\mathbf{l})} \mathbf{x}'^{(\mathbf{l})} \right) \right] \boldsymbol{\Delta} + \boldsymbol{\Sigma}_{\epsilon} \tag{9.36}$$

which can be contrasted with its univariate counterpart given by equation (9.9). In the remainder of this section the $r(k+1) \times r(k+1)$ matrix $Cov(\mathbf{z}^{(\mathbf{m})}) = \boldsymbol{\Sigma}_{\mathbf{z}} \otimes \left(\mathbf{x}^{(\mathbf{l})} \mathbf{x}'^{(\mathbf{l})} \right)$ will be denoted by $\boldsymbol{\Sigma}^{\otimes}_{\mathbf{z}}$.

$$Cov(\mathbf{z}^{(\mathbf{m})}) = \left[\begin{array}{cccc|cccc|c|cccc}
\sigma^2_{z_1} & x_1\sigma^2_{z_1} & \dots & x_k\sigma^2_{z_1} & \sigma_{z_1z_2} & x_1\sigma_{z_1z_2} & \dots & x_k\sigma_{z_1z_2} & \dots & \sigma_{z_1z_r} & x_1\sigma_{z_1z_r} & \dots & x_k\sigma_{z_1z_r} \\
x_1\sigma^2_{z_1} & x_1^2\sigma^2_{z_1} & \dots & x_1x_k\sigma^2_{z_1} & x_1\sigma_{z_1z_2} & x_1^2\sigma_{z_1z_2} & \dots & x_1x_k\sigma_{z_1z_2} & \dots & x_1\sigma_{z_1z_r} & x_1^2\sigma_{z_1z_r} & \dots & x_1x_k\sigma_{z_1z_r} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_k\sigma^2_{z_1} & x_1x_k\sigma^2_{z_1} & \dots & x_k^2\sigma^2_{z_1} & x_k\sigma_{z_1z_2} & x_1x_k\sigma_{z_1z_2} & \dots & x_k^2\sigma_{z_1z_2} & \dots & x_k\sigma_{z_1z_r} & x_1x_k\sigma_{z_1z_r} & \dots & x_k^2\sigma_{z_1z_r} \\
\hline
\sigma_{z_1z_2} & x_1\sigma_{z_1z_2} & \dots & x_k\sigma_{z_1z_2} & \sigma^2_{z_2} & x_1\sigma^2_{z_2} & \dots & x_k\sigma^2_{z_2} & \dots & \sigma_{z_2z_r} & x_1\sigma_{z_2z_r} & \dots & x_k\sigma_{z_2z_r} \\
x_1\sigma_{z_1z_2} & x_1^2\sigma_{z_1z_2} & \dots & x_1x_k\sigma_{z_1z_2} & x_1\sigma^2_{z_2} & x_1^2\sigma^2_{z_2} & \dots & x_1x_k\sigma^2_{z_2} & \dots & x_1\sigma_{z_2z_r} & x_1^2\sigma_{z_2z_r} & \dots & x_1x_k\sigma_{z_2z_r} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_k\sigma_{z_1z_2} & x_1x_k\sigma_{z_1z_2} & \dots & x_k^2\sigma_{z_1z_2} & x_k\sigma^2_{z_2} & x_1x_k\sigma^2_{z_2} & \dots & x_k^2\sigma^2_{z_2} & \dots & x_k\sigma_{z_2z_r} & x_1x_k\sigma_{z_2z_r} & \dots & x_k^2\sigma_{z_2z_r} \\
\hline
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
\hline
\sigma_{z_1z_r} & x_1\sigma_{z_1z_r} & \dots & x_k\sigma_{z_1z_r} & \sigma_{z_1z_2} & x_1\sigma_{z_1z_2} & \dots & x_k\sigma_{z_1z_2} & \dots & \sigma^2_{z_r} & x_1\sigma^2_{z_r} & \dots & x_k\sigma^2_{z_r} \\
x_1\sigma_{z_1z_r} & x_1^2\sigma_{z_1z_r} & \dots & x_1x_k\sigma_{z_1z_r} & x_1\sigma_{z_1z_2} & x_1^2\sigma_{z_1z_2} & \dots & x_1x_k\sigma_{z_1z_2} & \dots & x_1\sigma^2_{z_r} & x_1^2\sigma^2_{z_r} & \dots & x_1x_k\sigma^2_{z_r} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
x_k\sigma_{z_1z_r} & x_1x_k\sigma_{z_1z_r} & \dots & x_k^2\sigma_{z_1z_r} & x_k\sigma_{z_1z_2} & x_1x_k\sigma_{z_1z_2} & \dots & x_k^2\sigma_{z_1z_2} & \dots & x_k\sigma^2_{z_r} & x_1x_k\sigma^2_{z_r} & \dots & x_k^2\sigma^2_{z_r}
\end{array}\right]$$

Figure 9.10. The matrix $Cov(\mathbf{z}^{(\mathbf{m})})$. From: [105]

9.8.1 Estimation of the Mean and Variance Models, Multiple Response Case

The following section parallels the single response description given earlier (Section 9.5), and readers may want to contrast the material in this section against that other section.

By simply substituting $\boldsymbol{\Delta}, \boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}_\epsilon$ by their respective unbiased estimators, $\widehat{\boldsymbol{\Delta}}, \widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{\Sigma}}_\epsilon$ in equation (9.33), we obtain an unbiased estimator for $E(\mathbf{Y})$. To see this let:

$$\widehat{E}(\mathbf{Y}) = \widehat{\boldsymbol{\Theta}}'\mathbf{x}$$

Then, we have that:

$$E(\widehat{E}(\mathbf{Y})) = \boldsymbol{\Theta}'\mathbf{x} = E(\mathbf{Y}).$$

The estimator of the variance-covariance matrix of the response, $\widehat{Var}(\mathbf{Y})$, obtained from equation (9.36):

$$\widehat{Var}(\mathbf{Y}) = \widehat{\boldsymbol{\Delta}}'\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\widehat{\boldsymbol{\Delta}} + \widehat{\boldsymbol{\Sigma}}_\epsilon, \tag{9.37}$$

is not an unbiased estimator of $Var(\mathbf{Y})$. Similar steps can be followed as in the univariate case (previous section) to obtain an unbiased estimator. That is, to find an unbiased estimator of $Var(\mathbf{Y})$, find the expected value of the "naive" estimator in equation (9.37) and then, if possible, correct it by an unbiased estimator of the bias. To do this, let $\widehat{\boldsymbol{\delta}}_{.j}$ be the j^{th} column of $\widehat{\boldsymbol{\Delta}}$ and let $\widehat{\sigma}_{ij}$ be the $(ij)^{th}$ element of $Var(\mathbf{Y})$. We then have that:

$$\widehat{\sigma}_{ij} = \widehat{\boldsymbol{\delta}}'_{.i}\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\widehat{\boldsymbol{\delta}}_{.j} + \widehat{\sigma}_{\epsilon_{ij}}$$

Expectation can be taken yielding:

$$E(\widehat{\sigma}_{ij}) = \boldsymbol{\delta}'_{.i}\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\boldsymbol{\delta}_{.j} + tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\boldsymbol{\Sigma}_{\widehat{\delta}_{.i}\widehat{\delta}_{.j}}\right) + \sigma_{\epsilon_{ij}} \tag{9.38}$$

where $\boldsymbol{\Sigma}_{\widehat{\delta}_{.i}\widehat{\delta}_{.j}}$ is the cross-covariance matrix between the vectors $\widehat{\delta}_{.i}$ and $\widehat{\delta}_{.j}$.

From Press [129, pp. 233–234] it can be seen that this cross-covariance is:

$$\boldsymbol{\Sigma}_{\widehat{\delta}_{.i}\widehat{\delta}_{.j}} = \sigma_{\epsilon_{ij}}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}$$

where $\mathbf{X}_{\boldsymbol{\Delta}}$ is a matrix formed by the columns of the design matrix $\mathbf{X}$ corresponding to the regressors in $\mathbf{z}^{(\mathbf{m})}$. Hence $(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}$ is just the scaled covariance matrix of any of the columns of matrix $\widehat{\boldsymbol{\Delta}}$. These two matrices are better illustrated numerically in the example below. Note that every column of $\widehat{\boldsymbol{\Delta}}$ has the same "scaled" covariance matrix, they just differ by a constant, i.e. $\sigma_{\epsilon_{ij}}$.

Substituting $\boldsymbol{\Sigma}_{\widehat{\delta}_{.i}\widehat{\delta}_{.j}}$ in (9.38) we get:

$$E(\widehat{\sigma}_{ij}) = \boldsymbol{\delta}'_{.i}\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\boldsymbol{\delta}_{.j} + \sigma_{\epsilon_{ij}} tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right) + \sigma_{\epsilon_{ij}} \tag{9.39}$$

Since the argument of the tr operator does not depend on i or j the result presented in equation (9.39) can be extended to matrix form in the following way:

$$E(\widehat{Var}(\mathbf{Y})) = \boldsymbol{\Delta}'\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\boldsymbol{\Delta} + \left(1 + tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right)\right)\boldsymbol{\Sigma}_{\epsilon}$$

Finally, an unbiased estimator of $Var(\mathbf{Y})$ is given by:

$$\widehat{Var}(\mathbf{Y}) = \widehat{\boldsymbol{\Delta}}'\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\widehat{\boldsymbol{\Delta}} + \left(1 - tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right)\right)\widehat{\boldsymbol{\Sigma}}_{\epsilon} \tag{9.40}$$

In this case, the term $tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right)$ is the bias correction factor (compare with equation (9.16) for the univariate case).

Notice that if $tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right) > 1$, then the bias correction factor will be negative and if the elements in this matrix are sufficiently large, then the unbiased estimator will be given by a non-positive definite matrix. An equivalent problem is also present in the univariate case where a similar correction is made (see Section 9.5).

The possible non-positive definiteness of $\widehat{Var}(\mathbf{Y})$ arises from the fact that equation (9.36) only takes into account the variance due to the noise factors and the residual noise, not the variance in the estimation of $\boldsymbol{\Delta}$, a component estimated by the term $tr\left(\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}(\mathbf{X}'_{\boldsymbol{\Delta}}\mathbf{X}_{\boldsymbol{\Delta}})^{-1}\right)\widehat{\boldsymbol{\Sigma}}_{\epsilon}$

9.8.2 Solving the Multiple Response Robust Parameter Design Problem

Similarly as in the univariate RPD problem, it may be of interest in some applications to determine under which conditions the variability due to the noise factors can be made equal to zero. Ideally, the user would desire to make all the elements in $\boldsymbol{\Delta}'\boldsymbol{\Sigma}_{\mathbf{z}}^{\otimes}\boldsymbol{\Delta}$ equal to zero. Since this matrix is at least positive semi-definite, to achieve such a goal it is sufficient to make all the diagonal elements zero. That is, for this type of matrices, since $|a_{ij}| \leq \sqrt{a_{ii}a_{jj}}$, if $a_{ii} = 0$, then $a_{ij} = 0\ \forall j$. Miro et al. [105] show how a zero covariance matrix for the responses can only be obtained by making all the individual response variances zero and that this requires the availability of rq controllable factors,

which could be a large number in many practical situations. Even if such a number of controllable factors exist, the inclusion of all of them in a single experimental design will probably represent too many experiments. In addition, the solution $\mathbf{x}^*$ may lie outside the experimental region or could imply levels of the controllable factors that are unattainable in practice.

Notice that even when $\mathbf{\Sigma_z}$ and $\mathbf{\Sigma}_\epsilon$ are diagonal matrices, i.e. when the noise factors ($\mathbf{z}$) and the residual noise ($\boldsymbol{\epsilon}$) are each formed by independent random vectors, the correlation between two responses will not be zero if they happen to interact with at least one common noise variable. Therefore, for the multivariate case, an additional *covariance structure* will be generated by the noise factors.

Instead of looking at an unconstrained solution, we should seek a constrained solution to the multivariate RPD problem. In analogy with the univariate case, this involves constrained minimization of some scalar function of the covariance matrix.

Scalar functions of matrices are common for the generation of optimal designs according to various criteria (see Section 5.7). D-optimality minimizes the determinant of the scaled covariance matrix (sometimes called the generalized variance) while A-optimality minimizes the trace. Press (1982) mentions that other common scalar measures of internal scatter are $(tr\ Var(\mathbf{Y})^2)^{1/2}$ and the difference between the highest and smallest eigenvalue of $Var(\mathbf{Y})$. As it is well-known, the determinant is proportional to the volume of a joint confidence ellipsoid for $\mathbf{Y}$ [76].

With the exception of the determinant, which is discussed later in this section, the minimization of either of the aforementioned scalar objectives would be significantly impacted by the scaling of $Var(\mathbf{Y})$. From equation (9.36) we can see that a proper scaled covariance matrix of the responses can be obtained by scaling the regression coefficients in $\mathbf{\Delta}$, which in turn can be obtained by scaling the values of the responses used to fit the regression model. Using L_p norms is a common technique used in multi-criteria optimization to tackle this kind of problems, see for example Steuer [147] for a discussion on this issue.

We can finally state the optimization problem in the following way:

$$\begin{aligned} min \quad & f\left(Var(\mathbf{Y})\right) \\ subject\ to: & \\ & \mathbf{l} \leq E(\mathbf{Y}) \leq \mathbf{u} \\ & \mathbf{x} \in X \end{aligned}$$

where $\mathbf{u}$ and $\mathbf{l}$ are $q \times 1$ vectors containing upper and lower bounds for the response, $f(\cdot)$ is a suitable scalar function and $Var(\mathbf{Y})$ and $E(\mathbf{Y})$ are given by equations (9.36) and (9.33) respectively.

There has been considerable debate in the literature of "dual response systems" about what constitutes the best formulation, in particular, whether a MSE objective or a variance objective subject to constraints in the mean would work better. The MSE considers both the variance and the "bias" of the responses with respect to targets. The argument in favor of MSE is that by sacrificing some bias, a considerable reduction of variance is obtained. This argument is well-known to be true in estimation theory, but the RPD problem is no ordinary estimation problem. The MSE objective is even more non-linear (a quartic polynomial if the responses are quadratic), and simply minimizing the variance subject to *inequality* constraints in the mean response(s) has a similar behavior than minimizing the MSE, with the advantage of an easier to solve optimization problem.

An advantage of using the determinant is that *the optimization problem* defined previously is invariant to the scaling of the responses[15] (see Problem 6). Being invariant to scaling is a strong advantage since it avoids the possibility of having scale-dependent solutions. Therefore, using the determinant as the scalar function has the advantages that it does not requires any type of scaling, it considers the covariance structure in the responses, and it has the practical interpretation of being proportional to the joint confidence ellipsoid of the vector of responses. However, it requires the estimation of the complete covariance matrix as stated in equation (9.40).

In contrast, using the trace has the advantage that it only requires the diagonal elements of $Var(\mathbf{Y})$, which can be estimated from the univariate models. However, it does not consider the covariance structure and its interpretation depends on the normalizing method used to bring the individual variances to a common scale.

Example. Whey Protein Concentrates. Khuri & Cornell [76] report an experiment performed to investigate the effects of heating temperature (x_1), pH level (x_2), redox potential (x_3), sodium oxalate (x_4) and sodium lauryl sulfate (x_5) on foaming properties of whey protein concentrates. Measurements were made on three responses, the whipping time (Y_1), the maximum overrun (Y_2)

[15]The value of *determinant* itself evidently changes with scaling, but the functional dependence on $\boldsymbol{x}$ is not altered, hence the optimal solution obtained will not change depending on scaling.

and percent soluble protein (Y_3). Table 9.11 contains the experimental design and the multi-response data. Suppose that x_1 and x_3 are noise factors.

The $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{\Theta}}$ matrices were obtained after normalizing each response by its corresponding L_2 norm[16] variables. If we set to zero any non-significant parameter (avoiding non-hierarchical models), the corresponding matrices of parameter estimates are as follows:

$$\widehat{\mathbf{\Theta}} = \left[\begin{array}{l|rrr} & Y_1 & Y_2 & Y_3 \\ \hline Int. & 0.1692 & 0.1812 & 0.1813 \\ x_2 & 0.0068 & -0.0031 & 0.0145 \\ x_4 & -0.0068 & 0.0037 & 0.0022 \\ x_5 & -0.0030 & 0.0032 & 0.0032 \\ x_2x_4 & -0.0181 & 0 & 0 \\ x_2x_5 & 0 & 0 & 0 \\ x_4x_5 & -0.0125 & 0 & 0 \\ x_2^2 & 0 & -0.0022 & -0.0071 \\ x_4^2 & 0 & -0.0039 & -0.0019 \\ x_5^2 & 0 & 0 & 0.0040 \end{array}\right]$$

$$\widehat{\mathbf{\Delta}} = \left[\begin{array}{l|rrr} & Y_1 & Y_2 & Y_3 \\ \hline x_1 & 0.0460 & -0.0297 & -0.0159 \\ x_1x_2 & 0.0283 & -0.0042 & 0 \\ x_1x_4 & 0 & 0 & 0.004 \\ x_1x_5 & 0.0113 & -0.0062 & 0 \\ x_3 & -0.0159 & 0.0097 & 0.0046 \\ x_2x_3 & -0.0102 & 0 & 0 \\ x_3x_4 & 0 & 0 & 0 \\ x_3x_5 & -0.0091 & 0 & 0.0079 \end{array}\right]$$

The normalized estimated covariance matrix for the residuals of the fitted models is given by:

$$\widehat{\mathbf{\Sigma}}_\epsilon = 10^{-3} \begin{bmatrix} 3.2580 & -0.7132 & -1.3049 \\ -0.7132 & 0.5304 & 0.3697 \\ -1.3049 & 0.3697 & 0.6347 \end{bmatrix}$$

[16]That is, using $\mathbf{Y}_j/||\mathbf{Y}_j||,\ j = 1, \ldots, q$ to build up the response matrix $\mathbf{Y}$.

Table 9.11. Factor and Response Data for the Whey Protein Example. Source: Khuri and Cornell [76]

Controllable Factors					**Responses**		
x_1	x_2	x_3	x_4	x_5	Y_1	Y_2	Y_3
−1	−1	−1	−1	1	4.75	1082	81.4
1	−1	−1	−1	−1	4.00	824	69.6
−1	1	−1	−1	−1	5.00	953	105.0
1	1	−1	−1	1	9.50	759	81.2
−1	−1	1	−1	−1	4.00	1163	80.8
1	−1	1	−1	1	5.00	839	76.3
−1	1	1	−1	1	3.00	1343	103.0
1	1	1	−1	−1	7.00	736	76.9
−1	−1	−1	1	−1	5.25	1027	87.2
1	−1	−1	1	1	5.00	836	74.0
−1	1	−1	1	1	3.00	1272	98.5
1	1	−1	1	−1	6.50	825	94.1
−1	−1	1	1	1	3.25	1363	95.9
1	−1	1	1	−1	5.00	855	76.8
−1	1	1	1	−1	2.75	1284	100.0
1	1	1	1	1	5.00	851	104.0
−2	0	0	0	0	3.75	1283	100.0
2	0	0	0	0	11.00	651	50.5
0	−2	0	0	0	4.50	1217	71.2
0	2	0	0	0	4.00	982	101.0
0	0	−2	0	0	5.00	884	85.8
0	0	2	0	0	3.75	1147	103.0
0	0	0	−2	0	3.75	1081	104.0
0	0	0	2	0	4.75	1036	89.4
0	0	0	0	−2	4.00	1213	105.0
0	0	0	0	2	3.50	1103	113.0
0	0	0	0	0	3.50	1179	104.0
0	0	0	0	0	3.50	1183	107.0
0	0	0	0	0	4.00	1120	104.0
0	0	0	0	0	3.50	1180	101.0
0	0	0	0	0	3.00	1195	103.0

Suppose the covariance matrix of the two noise factors is known and equal to:

$$\mathbf{\Sigma_z} = \begin{bmatrix} 1.00 & -0.25 \\ -0.25 & 1.00 \end{bmatrix}$$

The bias correction factor is then given by:

$$1 - tr\left(\mathbf{\Sigma_z^{\otimes}}(\mathbf{X'_\Delta X_\Delta})^{-1}\right) = \frac{11}{12} - \frac{1}{8}\left(x_2^2 + x_4^2 + x_5^2\right)$$

The estimate of the covariance matrix of the responses is given by:

$$\widehat{Var}(\mathbf{Y}) =$$
$$\widehat{\mathbf{\Delta}}' \left[\begin{bmatrix} 1.00 & -0.25 \\ -0.25 & 1.00 \end{bmatrix} \otimes \left(\mathbf{x}^{(1)}\mathbf{x}'^{(1)}\right) \right] \widehat{\mathbf{\Delta}} + \left(\tfrac{11}{12} - \tfrac{1}{8}\left(x_2^2 + x_4^2 + x_5^2\right)\right) \widehat{\mathbf{\Sigma}}_\epsilon =$$

$$10^{-3} \begin{bmatrix} 2.73 + 3.39x_2 \\ +1.63x_5 + 1.05x_2^2 \\ +1.01x_2x_5 + 0.26x_5^2 & -1.75 - 1.29x_2 \\ & -0.83x_5 - 0.13x_2^2 \\ & -0.25x_2x_5 - 0.084x_5^2 & -0.92 + 0.20x_4 \\ & & -0.48x_5 - 0.56x_2 \\ & & +0.12x_2x_4 - 0.13x_2x_5 \\ & & +0.054x_4x_5 - 0.094x_5^2 \\ -1.75 - 1.29x_2 & 1.12 + 0.27x_2 & 0.59 - 0.13x_4 \\ -0.83x_5 - 0.13x_2^2 & +0.40x_5 + 0.018x_2^2 & +0.24x_5 + 0.07x_2 \\ -0.25x_2x_5 - 0.084x_5^2 & +0.052x_2x_5 + 0.038x_5^2 & -0.017x_2x_4 + 0.008x_2x_5 \\ & & -0.025x_4x_5 + 0.012x_5^2 \\ -0.92 + 0.20x_4 & 0.59 - 0.13x_4 & \\ -0.48x_5 - 0.56x_2 & +0.24x_5 + 0.07x_2 & 0.31 - 0.14x_4 \\ +0.12x_2x_4 - 0.13x_2x_5 & -0.017x_2x_4 + 0.008x_2x_5 & +0.13x_5 + 0.016x_4^2 \\ +0.054x_4x_5 - 0.094x_5^2 & -0.025x_4x_5 + 0.012x_5^2 & -0.016x_4x_5 + 0.062x_5^2 \end{bmatrix}$$

$$+10^{-3}\left(\tfrac{11}{12} - \tfrac{1}{8}\left(x_2^2 + x_4^2 + x_5^2\right)\right) \begin{bmatrix} 3.2580 & -0.7132 & -1.3049 \\ -0.7132 & 0.5304 & 0.3697 \\ -1.3049 & 0.3697 & 0.6347 \end{bmatrix} \tag{9.41}$$

Observe how there is considerable control over the covariance matrix, since the controllable factors are present in all the elements of the matrix.

Suppose the bounds for the expected responses are:

$$\mathbf{l} = [-\infty \quad 800 \quad 100]$$

$$\mathbf{u} = [5.0 \quad 1100 \quad \infty]$$

That is, the expected value of the first response is restricted to be less that 5.0 minutes, the expected second response between 800% and 1100% and the the

third response should be higher than 100%. Suppose we choose to conduct the optimization over a spherical region of radius 2 for the controllable factors, thus we add the constraint:

$$\sqrt{x_2^2 + x_4^2 + x_5^2} \leq 2$$

From Table 9.12 we note that the results using the trace are substantially different to the ones obtained with the determinant. In fact, the corresponding $\mathbf{x}^*$ points are separated by a distance of 0.91 coded units which is quite significant. The values of the objectives are also substantially different. The trace of the covariance matrix obtained when the determinant is minimized is about 58% larger than the one obtained when the trace is minimized. Similarly, the determinant obtained when the trace is minimized is about 18% larger than when the determinant is minimized.

Furthermore, notice that the solution obtained using the trace and the variances of the first and second responses are significantly close to each other. Therefore, using the trace as objective may not capture the overall variability of the vector of responses and could be affected by individual responses, especially if more than one of them is minimized close to the same point. This underscores the importance of considering the complete covariance matrix instead of the individual variances only, and demonstrates the benefits of the proposed multivariate approach over using the univariate approach q times in parallel. ■.

9.8.3 Loss Function Approach to Multiresponse Robust Parameter Design

Romano et al. [136] proposed a less general model for multivariate RPD than (9.29) in which there is no correlation in the errors between different responses. The correlation between the fitted responses, if any, originates only from the common regressors. These authors propose a loss function objective function, following earlier work in multiresponse RPD by Pignatiello [127] and Tsui [152]. The loss function these authors propose is

$$L(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{T})'\boldsymbol{C}(\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{T})$$

where $\boldsymbol{C}$ is a squared matrix that models quadratic off target costs, and $\boldsymbol{T}$ is a vector of targets for the responses. The expected value of this loss function is

$$E[L(\boldsymbol{x}, \boldsymbol{z})] = E(\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{T})'\boldsymbol{C}E(\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{T}) + tr(\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{y}}) \qquad (9.42)$$

Table 9.12. Comparison of Scalar Objectives for Whey Protein Example. Source: [105]

			Value of the other Functions at Optimality				
Scalar fun.	$\mathbf{x}^* = [x_2^*\ x_4^*\ x_5^*]$	$\widehat{E}(\mathbf{Y}^*)$	$\lvert\widehat{Var}(\mathbf{Y})\rvert$	$tr(\widehat{Var}(\mathbf{Y}))$	$\widehat{Var}(Y_{1,u})$	$\widehat{Var}(Y_{2,u})$	$\widehat{Var}(Y_{3,u})$
$tr(\widehat{Var}(\mathbf{Y}))$	[0.17 0.23 −1.98]'	[4.9 1036.8 100]'	0.38e–10	0.0024	1.12	2.12e4	103.8
$\lvert\widehat{Var}(\mathbf{Y})\rvert$	[0.59 0.97 −1.65]'	[5.0 1026.8 100]'	0.32e–10	0.0038	2.14	2.73e4	73.4
$\widehat{Var}(Y_1)$	[0.17 0.17 −1.98]'	[4.9 1035.9 100]'	0.40e–10	0.0024	1.11	2.11e4	106
$\widehat{Var}(Y_2)$	[0.17 0.12 −1.99]'	[4.9 1035.1 100]'	0.41e–10	0.0024	1.11	2.11e4	108
$\widehat{Var}(Y_3)$	[0.77 0.99 −1.55]'	[4.9 1021.4 100]'	0.37e–10	0.0046	2.69	2.96e4	72.0

where $\boldsymbol{\Sigma_y}$ denotes the covariance matrix of the responses. This loss function is evidently a sensible objective that in principle one could try to minimize subject to constraints in the controllable factors, expected responses, etc, *if* it is appropriately estimated. Unfortunately, the authors in reference [136] simply used plug-in estimates to estimate this expected loss (giving a biased estimator), and furthermore, neglected the variance transmitted by the noise factors to the responses simply arguing that they are constant during the DOE. This misses the main point in this chapter, in which *after* fitting the model, one should consider the noise factors as random variables and should then try to neutralize the variability transmitted by the noise factors.

9.9 Problems

1 Show that the expression (9.3) for assessing the statistical gains due to running an experiment in Split plot form, is true. (Hint: consider s^2 a weighted average of $MS_{\text{WP error}}$ and $MS_{\text{SP error}}$ and substitute in the left hand side of (9.3).)

2 Derive expression (9.20) using the law of conditional variance.

3 Corroborate that (9.23) is unbiased for (9.20) by taking expected value.

4 The following experimental data relates to a semiconductor experiment where the response is transistor gain [109]. It is desirable to hold the gain within the 200 ± 20 range.

			Response values for the noise factors:					
		z_1	-1.0	1.0	-1.0	1.0		
		z_2	-1.0	-1.0	1.0	1.0	Average	Std. Dev.
x_1	x_2	x_3						
-1	-1	-1	118.9	65.7	95.3	92.4	93.08	21.77
1	-1	-1	153.7	229.4	119.9	251.5	188.63	62.07
-1	1	-1	196.7	170.9	234.2	166.6	192.10	31.06
1	1	-1	211.1	245.7	241.0	252.6	237.60	18.30
-1	-1	1	145.2	132.2	167.1	137.9	145.60	15.29
1	-1	1	125.4	201.6	185.5	267.3	194.93	58.36
-1	1	1	283.0	251.1	263.4	190.4	246.98	39.94
1	1	1	184.2	279.5	247.2	259.2	242.53	41.11

Treating the outer array as replicates, do the following:

a) solve the RPD problem using the unbiased estimate of the variance

b) solve the RPD problem using the variance of the predicted response, $\widehat{Var}_{\boldsymbol{z},\widehat{\boldsymbol{\beta}}}(\widehat{y}(\boldsymbol{x},\boldsymbol{z}))$

c) do a ridge analysis on the variance response, using the variance of the predicted response, $\widehat{Var}_{\boldsymbol{z},\widehat{\boldsymbol{\beta}}}(\widehat{y}(\boldsymbol{x},\boldsymbol{z}))$.

5 Show that the $C(\tau)$ functions of two different designs D_1 and D_2 will never "cross", and that they can only coincide for the same DOE.

6 Let $\mathbf{N}$ be a q by q diagonal matrix containing the scaling or normalizing factors for the q responses. Let the subscripts u and s refer to the un-scaled and scaled versions of the associated variance matrices. Then show that $\left|\widehat{Var}_s(\mathbf{Y})\right| = |\mathbf{N}'\mathbf{N}|\left|\widehat{Var}_u(\mathbf{Y})\right|$. Thus, the determinant changes with scaling, but the optimal solution $\boldsymbol{x}^*$ obtained from minimizing $\left|\widehat{Var}_s(\mathbf{Y})\right|$ will be invariant to scaling as the functional dependence on $\boldsymbol{x}$ is not altered.

7 Repeat the Whey protein example assuming instead that x_2 and x_4 are the noise factors. Use the same noise factor covariance as in the text.

8 Find an unbiased estimator of the loss function (9.42). Consider both the noise factors and the parameter estimates as random variables when taking expected values.

Chapter 10

ROBUST OPTIMIZATION**

Science is made of mistakes which are useful to make, because they lead, little by little, to the truth.

—Jules Verne (1828–1905)

Something is always discarded when the results of experiment are trimmed down to fit formulas and equations. That something, much or little, which is thrown away has frequently been of scientific importance equal to what is retained in the mathematics.

—E.T. Bell (1883–1960)

10.1 Introduction

In this chapter[1] we discuss robustness from a more general perspective, that of model building in general, without specific discussion of noise factors. This notion of robustness in the sense of lack of sensitivity of an optimal solution to variations in the model has always been a key idea in mathematical modeling, and in particular, in mathematical programming. This differs from environmental variation in the sense of Taguchi, as discussed in the previous chapter. Thus, in this chapter, no "noise factors" are assumed to exist.

Practically since their inception, deterministic optimization techniques have been concerned with *sensitivity analysis*, that is, what is the effect on the optimal solution of a mathematical programming problem when some of the coefficients in the formulation changes. More recently, the idea of *robust*

[1] This chapter contains somewhat more advanced material and may be skipped on a first reading.

optimization has been promoted in the field of discrete optimization (see, for example, [82]). In a recent paper, Xu and Albin [160] have adopted the ideas in robust optimization to a RSM setting. The idea is to *find a solution which is as insensitive as possible, or robust, with respect to variations in the parameter estimates*. As it turns out, the Xu-Albin approach is related to techniques discussed in previous chapters, in particular, to the idea of constructing a confidence region on the optimal settings of a process (see Chapter 7). In contrast to the discussion on confidence regions, the focus here is on finding a *single* solution that a process engineer can use that is robust to uncertainty in the model parameter estimates.

In this chapter we present the "Minimax Deviation method" for robust optimization of Xu and Albin. This is a method that attempts to protect against sampling variability of the parameter estimates in the model, hence it is a frequentist method. We relate this method to confidence regions on the optimal settings (Chapter 7) and to some other proposals for process optimization from the area of Stochastic Programming.

A natural alternative to the Xu-Albin method is to employ a Bayesian approach in which the uncertainty in the model parameters, considered as random variables, is incorporated in the optimization. Such Bayesian approach to process optimization is presented in Part V of this book.

10.2 Minimax Deviation Method

Suppose we wish to minimize a polynomial model $f(\boldsymbol{x}, \boldsymbol{\theta})$ where the model form is assumed known, but the p parameters $\boldsymbol{\theta}$ unknown, and $\boldsymbol{x}$ represents the k controllable factors.

If we ignore the sampling variability, we simply optimize $f(\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$, that is, we simply optimize the fitted model where $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ represents the ordinary least squares estimates based on an $N \times p$ design matrix $\boldsymbol{X}$ which is written with columns that match the terms in the polynomial model.

In the following we repeat the argument in Xu and Albin [160]. Suppose $\boldsymbol{C}$ is the confidence region of the parameters $\boldsymbol{\theta}$, for some confidence level $1 - \alpha$ (see Section 7.3). Let $\eta(\boldsymbol{\theta}) = \min_{\boldsymbol{x} \in R} f(\boldsymbol{x}, \boldsymbol{\theta})$, i.e., the best $\boldsymbol{x}$ within its feasible region R, for some parameter vector $\boldsymbol{\theta} \in \boldsymbol{C}$. We are interested in operating conditions $\boldsymbol{x}$ such that $f(\boldsymbol{x}, \boldsymbol{\theta}) - \eta(\boldsymbol{\theta})$ is close to zero as the parameters $\boldsymbol{\theta}$ vary within their confidence region. That is, we want a solution which is

insensitive, or robust, with respect to variations in $\boldsymbol{\theta}$. The problem is how to achieve this. Xu and Albin propose the following minimax approach:

$$\min_{\boldsymbol{x}\in R}\max_{\boldsymbol{\theta}\in \boldsymbol{C}}\{f(\boldsymbol{x},\boldsymbol{\theta})-\boldsymbol{\eta}(\boldsymbol{\theta})\} \tag{10.1}$$

where $\eta(\boldsymbol{\theta}) = \min_{\boldsymbol{x}\in R} f(\boldsymbol{x},\boldsymbol{\theta})$. The deviation $f(\boldsymbol{x},\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta})$ is the loss incurred when choosing $\boldsymbol{x}$ if the true values of the parameters equals to $\boldsymbol{\theta}$. Thus, $\max_{\boldsymbol{\theta}\in \boldsymbol{C}}\{f(\boldsymbol{x},\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta})\}$ is the largest loss that can be incurred when we let $\boldsymbol{\theta}$ vary within its confidence region. Finally, by choosing $\min_{\boldsymbol{x}\in R}\max_{\boldsymbol{\theta}\in \boldsymbol{C}}\{f(\boldsymbol{x},\boldsymbol{\theta}) - \boldsymbol{\eta}(\boldsymbol{\theta})$ we are choosing the $\boldsymbol{x}$ such that the greatest loss is minimized, i.e., we find the operating conditions $\boldsymbol{x}$ such that the deviation of $f(\boldsymbol{x},\boldsymbol{\theta})$ from its optimal value $\eta(\boldsymbol{\theta})$ is minimized.

Xu and Albin point out how this minimax optimization problem can be written as

$$\begin{aligned}&\min \quad z\\ &\quad s.t. \quad f(\boldsymbol{x},\boldsymbol{\theta})-\eta(\boldsymbol{\theta})\le z, \quad \forall \boldsymbol{\theta}\in \boldsymbol{C}, \boldsymbol{x}\in R\end{aligned}$$

where z is a dummy variable. This is a semi-infinite mathematical programming problem. It can be converted in the following finite mathematical program:

$$\begin{aligned}&\min \quad z\\ &\quad s.t. \quad f(\boldsymbol{x},\boldsymbol{\theta}^{(i)})-\eta(\boldsymbol{\theta}^{(i)})\le z, \quad i=1,2,3,\ldots 2^p \ , \boldsymbol{x}\in R\end{aligned} \tag{10.2}$$

where the $\boldsymbol{\theta}^{(i)}$ are 2^p corner points of an approximated confidence region constructed as follows:

$$\boldsymbol{\theta}^{(i)} = \widehat{\boldsymbol{\theta}} + \Gamma^{-1}\boldsymbol{z}^{(-1)} \tag{10.3}$$

where

$$\Gamma = \frac{(\boldsymbol{X}'\boldsymbol{X})^{1/2}}{s\sqrt{p\ F_{p,N-p,\alpha}}},$$

the $\boldsymbol{Z} = [\boldsymbol{z}^{(i)}] = [(z_1, z_2, \ldots, z_p)]$ is a matrix that contains all the 2^p factorial combinations of a (2^p) design with $z_j = \pm 1$, and s is the square root of the usual MSE estimate. The points $\boldsymbol{\theta}^{(i)}$ are the corner points of a tilted rectangular approximation to the confidence region that better approximates the true confidence region as it considers the correlation between the parameter estimates. The resulting optimization problem in (10.2) is a nonlinear

optimization program if $f(\boldsymbol{x}, \boldsymbol{\theta})$ is nonlinear in $\boldsymbol{x}$. Just as in Section 7.3, it is convenient not to estimate the intercept in the polynomial model, so we really work with $p - 1$ parameters, i.e., the vector $\boldsymbol{\theta}$ is $(p - 1) \times 1$, where p is the number of parameters β_i in the polynomial model.

10.2.1 Relation with Confidence Regions on the Optimal Settings

In the PCD approach to finding a confidence region on $\boldsymbol{x}^*$, the optimal settings of a constrained response surface problem (see Section 7.3), the right hand side of the confidence interval for $\eta(\boldsymbol{\theta}) - z(\boldsymbol{x})'\boldsymbol{\theta}$ is given by

$$\begin{aligned} &\max_{\boldsymbol{\theta} \in \boldsymbol{C}} \min_{\boldsymbol{w} \in R} \{(z(\boldsymbol{w}) - z(\boldsymbol{x}))'\boldsymbol{\theta}\} \\ &\quad = -\max_{\boldsymbol{\theta} \in \boldsymbol{C}} \{z(\boldsymbol{x})'\boldsymbol{\theta} - \min_{\boldsymbol{w} \in R} z(\boldsymbol{w})'\boldsymbol{\theta}\}. \end{aligned} \tag{10.4}$$

Thus, if $f(\boldsymbol{x}, \boldsymbol{\theta}) = z(\boldsymbol{x})'\boldsymbol{\theta}$, i.e., if the model is linear in the parameters as we do assume, then the expression above is just the negative of the inner maximization in (10.1), namely, $-\max_{\boldsymbol{\theta} \in \boldsymbol{C}} \{f(\boldsymbol{x}, \boldsymbol{\theta}) - \eta(\boldsymbol{\theta})\}$. In the PCD method, we want to find all $\boldsymbol{x}$ such that (10.4) is negative, because if this is true we can reject $H_o : \eta(\boldsymbol{\theta}) - z(\boldsymbol{x})'\boldsymbol{\theta} = 0$ at $\boldsymbol{x}$ and therefore $\boldsymbol{x}$ is not in the confidence region (CR) for $\boldsymbol{x}^*$. Recall that the CR for $\boldsymbol{x}^*$ is made up of all $\boldsymbol{x}$ such that (10.4) is *greater* than zero.

In the robust optimization approach (10.1), we are trying in contrast to make the negative of (10.4) as small as possible over all $\boldsymbol{x} \in R$. This is equivalent to making (10.4) as *large* as possible for some $\boldsymbol{x} \in R$. Thus, in the Maximim deviation method, *we are trying to find the operating conditions* $\boldsymbol{x}$ *for which* $H_o : \eta(\boldsymbol{\theta}) - f(\boldsymbol{x}, \boldsymbol{\theta}) = 0$ *is hardest to reject*. This will be the most robust $\boldsymbol{x}$ with respect to variation in $\boldsymbol{\theta} \in \boldsymbol{C}$.

This gives a different interpretation of the robust optimization approach: we are trying to find operating conditions $\boldsymbol{x}$ such that the "false alarm" probability $P(\text{reject } H_o \mid H_o \text{ is true}) = P(\text{Type I error})$ is minimized. Here H_o simply says that the response value at $\boldsymbol{x}$ is indistinguishable from the optimum response value.

Example. We consider the example in Xu and Albin [160] who presented a simulated process where the true response, to be minimized, is given by

$$y = -2x_1 - 6x_2 - x_1x_2 + +0.5x_1^2 + x_2^2.$$

This function is to be minimized subject to the following constraints:

$$\begin{aligned} x_1 + x_2 &\leq 6 \\ x1 &\geq 0 \\ x2 &\geq 0 \end{aligned}$$

The minimum value of the true response subject to these constraints can be easily found to be $y = -19.6$ obtained at $x_1 = 2.8, x_2 = 3.2$. To investigate the performance of the Minimax deviation approach to robust optimization, Xu and Albin simulated the response surface y for one hundred DOEs each consisting of a 3^2 design. For each simulated response value, noise $\varepsilon \sim N(0, \sigma^2)$ was added. The simulated DOEs were repeated for 5 different values of σ: 0.5, 1, 2, 3, and 4.

For each simulated DOE, a full quadratic response was fitted. The fitted responses were then optimized by simply finding the maximum of the fitted response (what can be called the "Canonical" approach) and by using the Minimax deviation method. Finally, for each optimal solution $\boldsymbol{x}^*$, the value of the true response surface at $\boldsymbol{x}^*$ was saved.

Table 10.1 shows the estimated means, medians, and standard deviations of the response at the optimal settings obtained from both the canonical and the minimax deviation approaches ($\alpha = 0.05$ was used in the minimax approach). Similar results were shown by Xu and Albin [160]; we utilized the Matlab program `MinimaxDeviationEnhanced.m` (see below) to solve

Table 10.1. Mean, median, and std. deviation of the response surface obtained at the estimated operating conditions with the "canonical" method, which neglects the sampling variability, and the Minimax deviation method of Xu and Albin. The true minimum response surface value is -19.6. Results based on 100 replications of simulated 3^2 designs

	Canonical (standard) approach			Minimax Deviation approach		
σ	$\mu_{\boldsymbol{y}}$	median (y)	σ_y	μ_y	median (y)	σ_y
0.5	-18.5	-19.16	1.83	-18.37	-19.11	1.98
1	-15.43	-17.34	4.93	-15.62	-17.05	4.01
2	-8.18	-9.48	9.43	-11.14	-12.23	6.63
3	-6.56	-7.11	8.37	-8.9	-10.04	6.61
4	-4.94	-3.8	8.32	-7.69	-8.25	7.06

for the optimal $\boldsymbol{x}^*$ at each simulated DOE. As it can be seen from the Table, for little noise (small σ) the average response is close to its true value of -19.6. But as the noise level increases, the response values obtained at the estimated optimal settings get significantly worse for both methods. However, the deterioration in the quality of the response values is considerably less dramatic for the Minimax deviation method, indicating the *robustness* of the method to parameter estimates uncertainty created by a noisy process.

10.3 Relation with Stochastic Programming Methods in RSM

There have been proposals similar to the minimax deviation approach that we comment briefly in this section. In a recent paper, Diaz-Garcia et al. [47] propose to study a ridge-analysis problem using stochastic programming techniques. Assuming a second order polynomial model $\widehat{Y}(\boldsymbol{x}, \widehat{\boldsymbol{\beta}})$ has been fit, the problem they suggest to solve is

$$\min_{\boldsymbol{x}} \ \widehat{Y}(\boldsymbol{x}, \widehat{\boldsymbol{\beta}})$$

subject to

$$||\boldsymbol{x}||^2 \leq \rho^2$$

where

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1}) \quad \text{and} \quad \frac{n-p)\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p}$$

and $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma^2}$ are independent. This is a stochastic programming problem. The authors suggest to transform this problem into an equivalent deterministic problem in the sense that a solution to the deterministic problem is also a solution to the stochastic problem. One such equivalent deterministic problem is the minimax problem

$$\min_{\boldsymbol{x} \in R} \max_{\vartheta \in \Theta} \ E(Y(\boldsymbol{x}, \boldsymbol{\beta})$$

where $R = \{\boldsymbol{x} : ||\boldsymbol{x}||^2 \leq \rho^2\}$ and

$$\Theta = \left\{ \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} : \begin{array}{c} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{X'X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - p\widehat{\sigma}^2 F_{p,n-p,\alpha} \leq 0 \\ \text{and} \\ \frac{(n-p)\widehat{\sigma^2}}{\chi^2_{n-p,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-p)\widehat{\sigma^2}}{\chi^2_{n-p,\alpha/2}} \end{array} \right\}.$$

The solution suggested in [47] to this problem consists in solving two optimization problems in tandem. First, solve[2]

$$\max_{\boldsymbol{\vartheta}} \; Y(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{z}'(\boldsymbol{x})\boldsymbol{\beta}$$

subject to

$$\begin{aligned} &(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - p\widehat{\sigma}^2 F_{p,n-p,\alpha} \leq 0 \\ &\frac{(n-p)\widehat{\sigma}^2}{\chi^2_{n-p,1-\alpha/2}} - \sigma^2 \leq 0 \\ &\sigma^2 - \frac{(n-p)\widehat{\sigma}^2}{\chi^2_{n-p,\alpha/2}} \leq 0 \end{aligned} \quad .$$

Call the solution to this problem $\widehat{\boldsymbol{\vartheta}_1} = (\widehat{\boldsymbol{\beta}_1}', \widehat{\sigma}^2)'$. The second equivalent problem to solve is

$$\min_{\boldsymbol{x}} \; z'(\boldsymbol{x})\widehat{\boldsymbol{\beta}_1}$$

subject to

$$||\boldsymbol{x}||^2 \leq \rho^2.$$

This minimax approach is similar to the Xu-Albin proposal. Both perform an optimization over the confidence region of the parameter estimates. The main difference is that the Xu-Albin problem suggests to solve a stochastic programming problem for the minimax *deviation* from the true optimum response function, $\eta(\boldsymbol{\theta})$, whereas the minimax approach of Diaz-Garcia et al. does not consider robustness issues.

Other alternative stochastic programming approaches for the ridge analysis problem are discussed in [47].

10.4 Computer Implementation of the Minimax Deviation Method

The Minimax Deviation method requires the use of a computer to form the set of constraints (10.2) and to solve the different optimization problems involved. The Matlab program `MinimaxDeviationMethod.m` performs these tasks[3]. It takes as arguments the experimental design matrix (X), the vector of observed responses (y), the number of controllable factors (k), and the

[2]Although the authors do not suggest what value of $\boldsymbol{x}$ to use, one can use the solution to the standard ridge analysis problem.

[3]This and other programs described in this text are available from the author's personal web page.

overall (Bonferroni) α level (alpha). The design matrix is not in model form, and the program will automatically expand it to a full quadratic model (using a different model requires re-coding of the first few lines). The function returns the best operating conditions found using the Minimax deviation approach (bestxRobust) and using the classical or canonical approach (bestxCanonical). The function call is:

```
function[bestxRobust,bestxCanonical]
    =MinimaxDeviationEnhanced(X,y,k,alpha)
```

The program uses the `fmincon` nonlinear optimization routine to solve the different constrained minimization problems together with the `lhsdesign` routine to generate initial starting points for the optimizer (the optimization problem is in general a non-convex minimization problem). Therefore, Matlab's Optimization and Statistics Toolboxes are required.

In the example of the previous section, this program was used repeatedly from a driving program that simulated the DOEs and the responses. The example referred to the minimization of the response y subject to the constraints:

$$\begin{aligned} x_1 + x_2 &\leq 6 \\ x_1 &\geq 0 \\ x_2 &\geq 0 \end{aligned}$$

In the program, these need to be specified for the `fmincon` optimizer to consider them. The way to do this is to define the `Ac` matrix of constraint coefficients and the `bc` vector of constraint right hand sides, and lower and upper vectors of bounds on the x's. For the constraints in the example, these will be:

```
Lower=[0 0 -1e50];
Ac=[ 1 1 0];
bc=6;
```

(The third element of the Lower vector is the dummy variable z in the Xu and Albin formulation). Upper bounds on the factors can similarly be defined in the `Upper` vector.

10.5 Problems

1 Use the Matlab program `MinimaxDeviationEnhanced` to find robust operating conditions for the metal cutting experiment in Chapter 1. Suppose the tool life needs to be maximized subject to the constraints $-1 < x_1 < -0.5$ (in coded units). Assume a full quadratic model for the response.

2 Repeat the previous problem for the "grade A" response in Table 2.6.

3 Repeat the previous problem for the "grade B" response in Table 2.6.

PART V

BAYESIAN APPROACHES IN PROCESS OPTIMIZATION

Chapter 11

INTRODUCTION TO BAYESIAN INFERENCE

Probability is the most important concept in current science, especially as nobody has the slightest idea what it means.

—Bertrand Rusell (1872–1970)

11.1 Introduction

Reverend Thomas Bayes, a Presbyterian Minister who lived in England in the 18th century[1] wrote a manuscript on "inverse probability" related to the binomial distribution. This was published posthumously in 1763. Bayes' goal was to make probability inferences about the parameter of a binomial distribution. In 1774, Laplace stated what is now known as Bayes' theorem in general form, working independently.

Bayesian inference combines prior beliefs about model parameters with evidence from data using Bayes' theorem. There is a subjective interpretation of probability in this approach, compared to the "frequentist" approach in which the probability of an event is the limit of a ratio of frequencies of events. The main criticisms of Bayesian analysis have been 1) that it is not objective (a fact that has been debated for many years), and 2) that the required computations are difficult. The second criticism has been overcome to a large extent in the last 10–15 years due to advances in integration methods, particularly, Markov Chain Monte Carlo (MCMC) method. The object of this chapter is to

[1]Few biographical details are known about Bayes. for some of them, see Press [130].

present an introduction to statistical inference problems from a Bayesian point of view. This will lead us in the next chapter to Bayesian regression and its use in process optimization.

11.2 Basics of Bayesian Inference

11.2.1 Notation

In this part of the book, θ will denote unobserved quantities or population parameters of interest ($\boldsymbol{\theta}$ will denote a vector of k such quantities), y will denote observable quantities (data), which can be a vector ($\boldsymbol{y}$, of n components); $\tilde{y}$ denotes a future observation of the same nature as y, and $\boldsymbol{X}$ denotes an $n \times p$ matrix of explanatory variables or covariates. $\boldsymbol{X}$ contains the experimental design.

Likewise, $p(\cdot)$ will denote a continuous density function, and the notation $w|y$ denotes a conditional random variable w given y (the data). The notation $P(\cdot)$ will refer to the probability of some event defined over a sample space. Sometimes we will simply write "data" for all the data obtained from an experiment.

11.2.2 Goals of Bayesian Inference

The goal of Bayesian inference is to reach conclusions about a parameter θ or future observation $\tilde{y}$ using probability statements *conditional* on the data $\boldsymbol{y}$. For doing this, we use three probability densities:

- $p(\theta)$, the density of θ before observing any data. This is called the prior density of θ is user-defined and should reflect whatever the user believes or thinks the possible values of θ are (hence, it is subjective);
- $p(\theta|y)$, the density of θ after observing experimental data, this is called the *posterior density* of θ; and
- $p(\tilde{y}|y)$, *the posterior predictive density* of a future observation y. This is the density of a future, not yet observed response value y after we have observed previous $y's$ from an experiment.

An important difference between Bayesian and Classical Statistics is as follows. Bayesian inference considers *all* unknowns (parameters and future observations) as random variables. Classical (frequentist) statistical inference

considers population parameters as fixed, but data as random (due to sampling). As we discuss below, there is a tendency in Bayesian statistics to consider unobservable parameters as intermediate variables not of interest *per se*, and to consider inferences on observable quantities the ultimate objective of Statistics.

11.3 Bayes' Theorem for Events

Let us first look at Bayes' theorem in its most elementary form, that of simple events. If events A, B, etc. occur in some sample space $\mathcal{S}$, we have, from the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

These expressions are true if and only if

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

from which

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

This gives the essence of Bayes' theorem: if event B represents some additional information that becomes available, then $P(A|B)$ is the probability after this information becomes available, i.e., the *posterior* of A, and $P(A)$ is the probability before this information becomes available, i.e., the *prior* for A.

Suppose events A_i form a partition of $\mathcal{S}$, that is, $\bigcap_{\text{all}i} A_i = \mathcal{S};\ A_i \cap A_j = \oslash$ for all $i, j,\ i \neq j$. Then, the ("total") probability of B is given by

$$P(B) = \sum_{\text{all}j} P(B|A_j)P(A_j)$$

and therefore

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{\text{all}j} P(B|A_j)P(A_j)}. \tag{11.1}$$

Expression (11.1) is what Laplace referred to as the problem of finding the "inverse probability": given that the "effect" B is observed, find which of several potential "causes" was the true cause of the observed effect.

Evidently, there is nothing incorrect in Bayes' formula, as it is derived from the probability axioms. The debate concerns in the interpretation of the probabilities involved. Classical statistics regards each probability in the formula as the limit of the ratio of frequencies; Bayesian statistics regards each probability as a subjective measure.

11.4 Bayes' Theorem for Densities

The derivation of Bayes' theorem for densities parallels that of the previous section. If y denotes data and θ some parameter or vector of parameters, from the definition of conditional density, we have that

$$p(y|\theta) = \frac{p(\theta, y)}{p(\theta)}$$

and

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)}.$$

This implies the joint density is

$$p(\theta, y) = p(\theta)p(y|\theta) = p(y)p(\theta|y)$$

which implies

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{11.2}$$

where analogous to the total probability of an event

$$p(y) = \int_{\text{all } \theta} p(\theta)p(y|\theta)d\theta. \tag{11.3}$$

Equation (11.2) is Bayes' law for densities. The denominator is usually not computed since it is not a function of θ (which is integrated out in (11.3)) and only makes $p(\theta|y)$ integrate to one in (11.2). Therefore, the Bayesian Statistics literature uses the proportionality sign $\propto$ and Bayes' formula, in its most common form is:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{11.4}$$

In words, *the posterior is proportional to the product of the prior times the likelihood of the data*. Note that the posterior probabilities are therefore proportional to the likelihood, and the likelihood is a central concept in classical statistics, e.g., in maximum likelihood estimation. What we do is to modify the

likelihood according to our prior beliefs of the parameter. If the prior is very "flat", Bayesian inferences will be very close to likelihood inferences.

Bayes' formula provides a recursive mechanism for updating the posterior distribution that is very useful in applications where the observations are obtained sequentially. As each new observation is obtained, the posterior is updated treating the previous posterior as the prior, forming a chain:

$$p(\theta) \Rightarrow p(\theta|y_1) \Rightarrow p(\theta|y_1, y_2) \Rightarrow p(\theta|y_1, y_2, y_3) \Rightarrow \cdots$$

Evidently, the theorem assures that if more than one observation is obtained at a time it is possible to "jump" two or more steps in the chain above, with identical results, i.e., for example

$$p(\theta) \Rightarrow p(\theta|y_1, y_2)$$

will result in the same distribution that would be obtained if going from $p(\theta)$ to $p(\theta|y_1, y_2)$ via $p(\theta|y_1)$. The sequential application of Bayes' theorem is a central idea in some Engineering applications, for example, in Kalman filtering (see Meinhold and Singpurwalla [97]).

To make *inferences on an unobservable* θ, we can simply look at the posterior distribution $p(\theta|y)$. This provides a *complete* characterization of our state of knowledge about the parameter and it is recommended to report it, perhaps using a histogram from a simulation or graphs if a closed-form expression exists. Still, there are some applications where a single value or guess is needed about the unknown parameter, or an interval of possible values is desired or needed. Such a single value is the analog of a classical "point estimate". Two usual Bayesian choices are the mode of the posterior distribution:

$$\hat{\theta} = \arg\max p(\theta|y),$$

which evidently implies a maximization problem, and the mean of the posterior distribution

$$\hat{\theta} = E[p(\theta|y)],$$

which implies an integration. Each choice has different properties. For example, the mean of $p(\theta|y)$ minimizes the expected square error. The mode maximizes the expected utility function when there is a unit benefit if $\hat{\theta} = \theta$ and zero benefit if $\hat{\theta} \neq \theta$.

A Bayesian interval estimator of a parameter θ or *Bayesian credibility interval* is given by the interval (a, b) such that $P(a < \theta < b|y) = 1 - \alpha$, where $1 - \alpha$ is the "credibility" level. This value, contrary to the confidence level of a classical interval, *is a probability*, and gives an indication of how probable it is that the parameter is contained within the computed interval. This interpretation is easier to grasp than the long-run coverage interpretation of a classical confidence interval. Sometimes, a "Highest Posterior Density" (HPD) interval is desired, which is obtained from solving:

$$\min \; b - a \quad s.t. \quad \int_a^b p(\theta|y)d\theta = 1 - \alpha.$$

For symmetric unimodal distributions, this is always a symmetric interval around the mode. For a multimodal distribution, HPD intervals are harder to get.

To make inferences about a *future observation*, $\tilde{y}$, we compute the *posterior predictive density* as follows:

$$\begin{aligned} p(\tilde{y}|y) &= \int_{\text{all } \theta} p(\tilde{y}, \theta|y)d\theta \\ &= \int_{\text{all } \theta} p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int_{\text{all } \theta} p(\tilde{y}|\theta)p(\theta|y)d\theta \end{aligned} \tag{11.5}$$

where the last equality follows since $\tilde{y}$ and y are conditionally independent given θ, that is, the parameters, if known, summarize the data. Similarly as for unobservable parameters, we can use summarizing measures to provide single estimates or intervals on $\tilde{y}$. Again, simply looking at the whole posterior predictive distribution is the most complete approach.

It is important to contrast the predictive density with the classical approach of making predictions on $\tilde{y}$. The classical approach uses $p(\tilde{y}|\hat{\theta})$, the data density evaluated at the maximum likelihood or least squares estimator, to make predictions. Unlike (11.5), this distribution *does not account for the uncertainty in estimating* θ a crucial issue for process optimization, since, as mentioned before (Chapters 6, 7, 9 and 10), different parameter estimates will result in different optimal solutions[2].

[2]Classical statistics can give an idea of the variability of the optimal solution due to parameter estimates, by computing confidence regions on the optimum (see Chapter 7). These regions, while useful, have the classic interpretation of being the region where the optimum could lie in hypothetical *repeated* experiments.

The posterior predictive density is the basis of Bayesian *predictivism*, a school of thought we now comment on.

11.5 Predictivism

It is helpful to think about the posterior predictive density as the probability of a new observation $\tilde{y}$ *averaged* over all possible posterior values of the parameters the likelihood depends on. Thus, the predictive density is a weighted average, with weighs equal to the posterior probability of the parameter values.

This weighted average is useful in scientific inference (see Press [130]), and can be illustrated easily in the case there is a discrete number of alternative "theories" we wish to test. Suppose $\theta = 1$ means "Theory A is true" and $\theta = 0$ means "Theory B is true", and only these two theories are entertained to explain a phenomena. We collect measurements, and after observing the data we compute

$$p(\tilde{y}|data) = p(\tilde{y}|\theta = 1)p(\theta = 1|data) + p(\tilde{y}|\theta = 0)p(\theta = 0|data)$$

which is the predictive probability of a new observation $\tilde{y}$.

Predictivism is a school of thought in the philosophy of science that postulates that the value of a scientific theory is measured by its ability to predict some phenomena, regardless of its ability to provide a mechanism that explains it. This idea is particularly useful in applied science and engineering. Within Bayesian statistics, predictivism states that the important quantities are the *observable* ones, not the unobservable ones (parameters). The posterior predictive density is the means to make predictions about and test a hypothesis. The counterpart of "model diagnostics" in Bayesian Statistics are based on the posterior predictive densities. In this type of diagnostics, we compare simulated predicted $\tilde{y}$'s using the posterior predictive density versus the data, and see how similar they look. If the data is very different than the simulated responses, it is an indication our model fails to represent reality well. We now turn to how to perform such simulations.

11.6 Simulation of Posterior Quantities

It is surprisingly easy to compute posterior probabilities of functions of random variables. Given $p(\theta|y)$ and $p(\tilde{y}|y)$, we can obtain posterior probabilities for functions of θ or $\tilde{y}$ as complex as needed.

Table 11.1. Simulation of the coefficient of variation

Draw number	**Parameters** ($\boldsymbol{\theta}$)	CV
1	μ_1, σ_1	μ_1/σ_1
2	μ_2, σ_2	μ_2/σ_2
$\vdots$	$\vdots$	$\vdots$
m	μ_m, σ_m	μ_m/σ_m

Example. To find the posterior distribution of the coefficient of variation CV$= \mu/\sigma$ of a N(μ, σ^2) distribution, let $\boldsymbol{\theta} = (\mu, \sigma)'$. Given $p(\boldsymbol{\theta}|y)$ we can simulate instances of $\boldsymbol{\theta}$ as shown in Table 11.1.

To perform the simulation of the posterior of the parameters, a widely used "trick" if the joint posterior is difficult to get analytically[3], is to note that:

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y).$$

Thus an algorithm for the simulation of the posterior of the CV will look like this (see Table 11.2 and Figure 11.1 for a Matlab implementation[4]):

1 collect n observations and compute $\overline{y}$ and s^2;

2 simulate $\sigma^2|y$;

3 simulate $\mu|\sigma^2, y$;

(pairs of simulated values above give $\mu, \sigma^2|y$)

4 compute μ/σ;

5 goto 2 until we iterate N times.

In Figure 11.1, the values $\overline{y} = 100$ and $s^2 = 10$ were observed based on a sample of size $n = 5$. Note how $p(\mu|\sigma^2, y)$ is centered around $\overline{y} = 100$ since the prior distribution was "flat" (more on this later in this chapter). ■

[3]The joint posterior is not so difficult to obtain in this case, but we will use this simpler case for illustration of this approach.

[4]To be more precise, and as it will be seen later, the program generates $\sigma^2|y \sim Inv - \chi^2(n-1, s^2)$ (an inverse chi squared distribution) and $\mu|\sigma^2, y \sim N(\overline{y}, \sigma^2/n)$. This is based on noninformative priors.

Table 11.2. Matlab simulation program for the posterior distribution of $CV = \mu/\sigma$

```
function [coeff,theta,sigma2]
     =CV(noOfSimulations,ybar,n,s2);

for i=1:1:noOfSimulations
    x=chi2rnd(n-1);
    sigma2(i)=(n-1)*s2/x;
    theta(i)=normrnd(ybar,sqrt(sigma2(i)/n));
    coeff(i)=theta(i)/sqrt(sigma2(i)/n);
end;
```

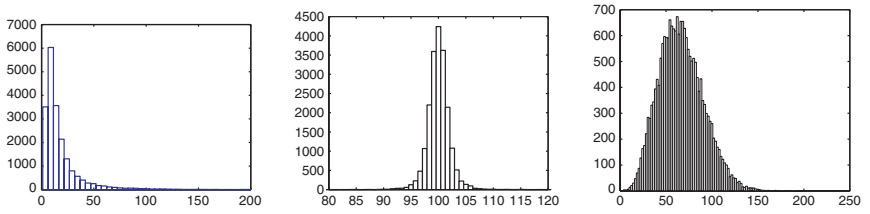


Figure 11.1. MATLAB simulation of the coefficient of variation of a Normal distribution, $N = 10,000$. Left: $p(\sigma^2|y)$; center: $p(\mu|\sigma^2, y)$; right: $p(CV|y)$

11.6.1 How to Simulate the Posterior Predictive Density

If the integral required to compute the posterior predictive density is hard to obtain, simulation is helpful. If $p(\tilde{y}|\theta)$ and $p(\theta|y)$ are available, simulation is an easy alternative. To simulate

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

we do the following:

1 simulate a value of θ from $p(\theta|y)$;

2 simulate a value of $\tilde{y}$ from $p(\tilde{y}|\theta)$;

3 Go to 1 unless N iterations are reached. A histogram of the N $\tilde{y}$'s characterizes $\tilde{y}|y$.

11.7 Choice of Prior Distribution

The most common and valid criticism in Bayesian Statistics is the question: "Where did the prior come from?" Any prior needs to be justified in practice. Three important choices of priors are:

1 Conjugate priors

2 Non-Conjugate priors

3 Non-informative priors (these are non-conjugate as well).

Conjugate priors. If $\mathcal{F}$ is a class of sampling distributions $p(y|\theta)$ and $\mathcal{P}$ is a class of prior distributions for θ, $p(\theta)$, then $\mathcal{P}$ is said to be *conjugate* for $\mathcal{F}$ if:

$$p(\theta|y) \in \mathcal{P} \ \ \forall \ \ p(y|\theta) \in \mathcal{F}, p(\theta) \in \mathcal{P}.$$

In words, this means that the prior and the posterior distributions of the parameter have the same form (with different parameters), so conjugacy is a closure property. The main merit of conjugate priors is that it simplifies computations, particularly in sequential applications of Bayes' theorem. With these distributions, the integral we need to compute for the posterior has a familiar form, hence the computational advantage. However, in many applications "tuning" a conjugate prior to reflect the knowledge of the user is a difficult problem, or the conjugate priors may not be able to reflect this knowledge. While there is some literature on "elicitation of priors" (see e.g., Kadane et al. [73] for the case of a regression model, which is our main emphasis in the next chapter), elicitation has had little impact on statistical practice.

Non-conjugate priors do not have the closure property of conjugate priors; they result in posteriors that have a different parametric form than the prior. Until recently, they were not discussed frequently in the literature, given the hard integrals involved. With the advent of Markov Chain Monte Carlo (MCMC) methods in the last decade, these priors have been use more often as they could be chosen to better reflect prior knowledge of the parameters.

Non-informative priors are non-conjugate, and try to reflect a situation where there is a complete lack of knowledge about a parameter. Therefore, they are called "objective priors" by some authors, and the resulting analysis is called objective Bayesian analysis. In applied problems, complete *a priori*

ignorance hardly exists in an experiment, so a non-informative prior should in practice be regarded as an approximation to a situation where *little* is known *a priori* [28].

We will consider determining non-informative priors for *location* and *scale* parameters of a distribution, so we need the following definitions.

Let $p(y)$ be any pdf. A *location-scale* family of density functions has the form:

$$\frac{1}{\sigma}\, p\left(\frac{y-\mu}{\sigma}\right)$$

where $-\infty < \mu < \infty$ is a location parameter and $\sigma > 0$ s a scale parameter. That is, μ shifts the location of the distribution on the y axis and σ stretches (contracts) the graph of $p(y)$ if $\sigma > 1$ ($\sigma < 1$). In either case, changing these parameters does not change the shape of the distribution. Examples of location-scale distributions are the normal distribution and the double exponential. If $\sigma = 1$, then a density of the form $p(y - \mu)$ is called a *location density*, and if $\mu = 0$ a density of the form $p(y/\sigma)/\sigma$ is called a *scale density*.

11.7.1 Non-informative Priors

It is tempting for a beginner in Bayesian Statistics to think of the Uniform distribution as an ideal candidate for a noninformative prior. It is important to realize that a uniform distribution over a finite range is informative in the sense that values of the parameter are excluded (if the prior is zero over some range, the posterior will be zero over that range). Such a prior was used by Bayes.

Laplace introduced the *principle of insufficient reason* for which he implied that in the absence of any information about a parameter, *all* values should be equally likely. Jeffreys generalized this reasoning into what is called the *invariance principle*. We will focus on invariance as a principle of finding non-informative priors, but readers should be aware that there have been many other approaches put forward to define a non-informative prior (see Kass and Wasserman [74] for an excellent review). In particular, there is a lot of debate about what "non-informative" means, and this has resulted in the agreement that the non-informative priors used are more for convenience than as a true description of lack of information. They should be used as "reference priors" in the sense of being a default choice that makes sense in the situation when one knows little about a parameter. Unfortunately, some Bayesian software (e.g., Winbugs) do not allow non-informative priors.

A summary of the non-informative priors we will use is as follows:

1 For parameters θ defined over a finite range of possible values $R \subset \mathbb{R}$, define the prior of θ to be uniform in R. An example of this was proposed by Bayes himself, who used a uniform (0, 1) on the binomial proportion parameter p;

2 For parameters θ defined over all reals $\mathbb{R}$, use a Uniform $(-\infty, \infty)$ distribution as prior;

3 For parameters θ defined over the positive real line $\mathbb{R}^+ \subset \mathbb{R}$, define a prior for $\log \theta$ to be Uniform in $(-\infty, \infty)$;

4 When trying to setup a non-informative prior in multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'$, apply the criteria 1–3 above to each parameter individually. This means that the parameters are independent *a priori*.

Evidently, only the first case above gives a proper prior distribution. Cases 2 and 3 lead to improper prior distributions, that is, density functions which do not have a finite integral. These improper priors can lead to proper posteriors in important cases as we will show shortly. However, proper posteriors will not always result, so one should always check that the posterior is proper when using improper priors.

The important thing in practice, however, is that *a non-informative prior should be flat where the likelihood is non-negligible*. See Figure 11.2, where a "flat" but proper prior is set on a parameter where the likelihood of the data given the parameter is non-negligible. Since we cannot know the location of the likelihood before the data is collected, we will not be able to use such a flat, proper prior, because we will not know where to locate it *a priori*. Instead, we use the improper priors in cases 1 and 2, which evidently are "everywhere uniform" over the possible values of either θ (case 1) or $\log\ \theta$ (case 2). The improper prior should be seen as a useful device to approximate the ideal case of Figure 11.2.

We now justify the choice of improper priors (cases 2 and 3 above) in more mathematical detail.

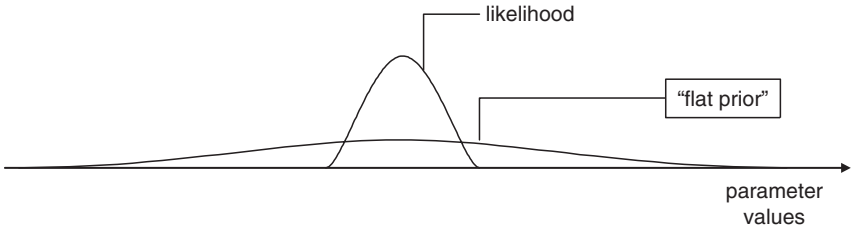


Figure 11.2. Hypothetical likelihood for a parameter θ defined over all reals, and a flat but proper prior density

11.7.2 Derivation of Invariant Non-informative Priors**

Location parameter.- Suppose[5] we observe a random variable

$$Y = X + c$$

where c is any real constant and where X is a random variable with a location density with location parameter θ, with possible values over *all* reals. Then Y is a location density with location parameter $\eta = \theta + c$ that can also take values over all reals. If we wish to setup a non-informative prior probability density on θ because we have complete ignorance of its true value, we should also setup a prior that should be equally non-informative for η. How to quantify "equally non-informative" has been debated for years. Jeffreys' invariance argument says that if we do not know anything about θ, since the origin or coordinates is arbitrary (e.g., we could have chosen Kelvin or Celsius degrees to measure a temperature) then whatever probability we assign to the event $\theta \in A$ where A is an arbitrary interval over $\mathbb{R}$, we should assign the same probability for $\eta \in A$. That is, we want to set a prior in either θ or η such that, for any interval A:

$$P(\theta \in A) = P(\eta \in A).$$

Figure 11.3 shows a case where the prior probability assignment does not obey this criterion. Note how

$$P(\eta \in A) = P(\theta + c \in A) = P(\theta \in (A - c))$$

where $A - c$ means we subtract c from all points in A. Therefore, for a non-informative prior we must have that

$$P(\theta \in A) = P(\theta \in (A - c)).$$

[5]This section contains somewhat more advanced material and may be skipped on a first reading.

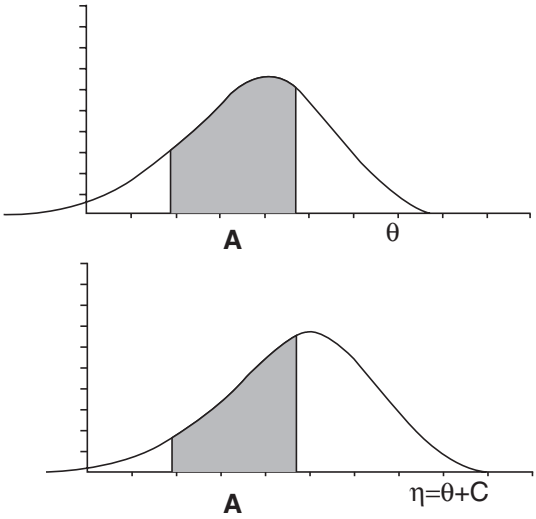


Figure 11.3. A prior on a location parameter θ which is not invariant with respect to a shift in location, i.e., $P(\theta \in A) \neq P(\eta \in A)$

In terms of the densities involved, the invariance principle requires that the priors satisfy:

$$\int_A p(\theta)d\theta = \int_{A-c} p(\theta)d\theta. \tag{11.6}$$

Applying the change of variable theorem[6]

$$\int_{A-c} p(\theta)d\theta = \int_A p(\theta - c)d\theta$$

so we have, from (11.6), that

$$\int_A p(\theta)d\theta = \int_A p(\theta - c)d\theta.$$

This can be true only if

$$p(\theta) = p(\theta - c) \ \forall \ \theta. \tag{11.7}$$

In particular, if $\theta = c$, we get $p(c) = p(0)$, as in Figure 11.4. This can only be true if $p(\theta)$ is constant for *all* $\theta \in \mathbb{R}$. Therefore, *the noninformative prior density for a location parameter* $\theta \in \mathbb{R}$ *is*

$$p(\theta) \propto \text{ constant over } (-\infty, \infty).$$

Evidently, this results in an improper prior.

[6]See Appendix D.

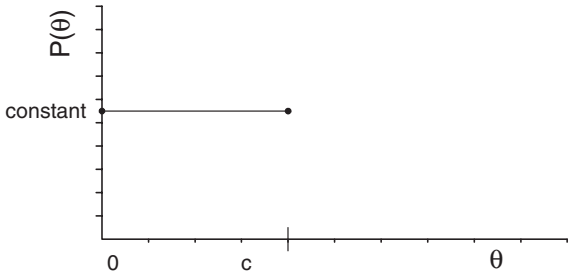


Figure 11.4. Illustration of expression (11.7) for $\theta = c$

Scale parameters. If someone is totally ignorant on a scale parameter, then she should set the probability of the parameter being in the interval (1,10), for example, equal to the probability of the parameter being in the (10,100) interval. This implies a flat prior in the logarithm of the parameter. More in detail, suppose we observe

$$Y = cX \;\; (c > 0)$$

where X is a random variable with scale density with parameter $\sigma > 0$. Applying the transformation of random variables theorem (see Appendix D) we have that $Y = cX$ implies $X = Y/c = w(Y)$, so $d/dy\; w(y) = 1/c$, and

$$p_Y(y) = p_X(x)\left|\frac{1}{c}\right| = \sigma^{-1} p_X\left(\frac{y}{c\sigma}\right)\frac{1}{c} \doteq \frac{1}{\eta} p_X\left(\frac{y}{\eta}\right)$$

which is also a scale density with scalar parameter $\eta = c\sigma > 0$.

Since the scale of measurements (e.g. kilograms or grams) is arbitrary, a non-informative prior should not change with changes in the scale (invariance principle). That is, the following probability statement should hold:

$$P(\sigma \in A) = P(\eta \in A) \;\forall\; A \subset \mathbb{R}.$$

In words, in the absence of any prior knowledge about the scale parameter, the probability of the parameter being inside any interval should be the same as the probability of a transformed scale parameter being inside the same interval. Since $\eta = c\sigma$ we must have

$$P(\eta \in A) = P(\sigma \in A/c) = P(\sigma \in A) \;\forall\; c > 0$$

where the division means to divide each point in A by c. Following a similar argument as in the location case, we want to set the prior density such that

$$\int_A p(\sigma)d\sigma = \int_{A/c} p(\sigma)d\sigma.$$

Applying the change of variable theorem, we get

$$\int_{A/c} p(\sigma)d\sigma = \int_A p\left(\frac{\sigma}{c}\right)\frac{1}{c}d\sigma,$$

which can be true only if

$$p(\sigma) = \frac{1}{c}p\left(\frac{\sigma}{c}\right) \ \forall\ \sigma > 0.$$

In particular, if we choose $\sigma = c$, we get

$$p(\sigma) = \frac{1}{\sigma}p(1).$$

Therefore, *the non-informative prior density for a scale parameter* $\sigma > 0$ *is*

$$p(\sigma) \propto \frac{1}{\sigma}.$$

This prior will be invariant to changes in scale. It is also improper, but for important cases (although not always) the posterior is proper.

Example. Illustration of change of scale on prior. Suppose $A = (1, 2)$ and consider other intervals A/c. Then the non-informative prior on a scale parameter σ looks like Figure 11.5, where we have illustrated the probability $P(\sigma \in A/c)$ for $c = 0.5$, 1, and 2. For $c = 1$ we get $P(\sigma \in A)$, but for all other cases we get $P(\eta \in A)$ since $\eta = c\sigma$. All the shaded areas are equal to ln(2) numerically. This implies that $P(\sigma \in A) = P(\eta \in A)$, as required. The lower part of the figure illustrates how each of the areas in the $p(\sigma)$ density graph map into the areas in the log σ scale, where the prior is improper uniform. ■

It should be pointed out that the non-informative priors derived in this section are not unique. Since they are improper integrals, any multiple of them will be equally acceptable. This does not reduce the utility of these priors as approximations to the situation where there is little knowledge about a parameter, provided the resulting posterior is proper.

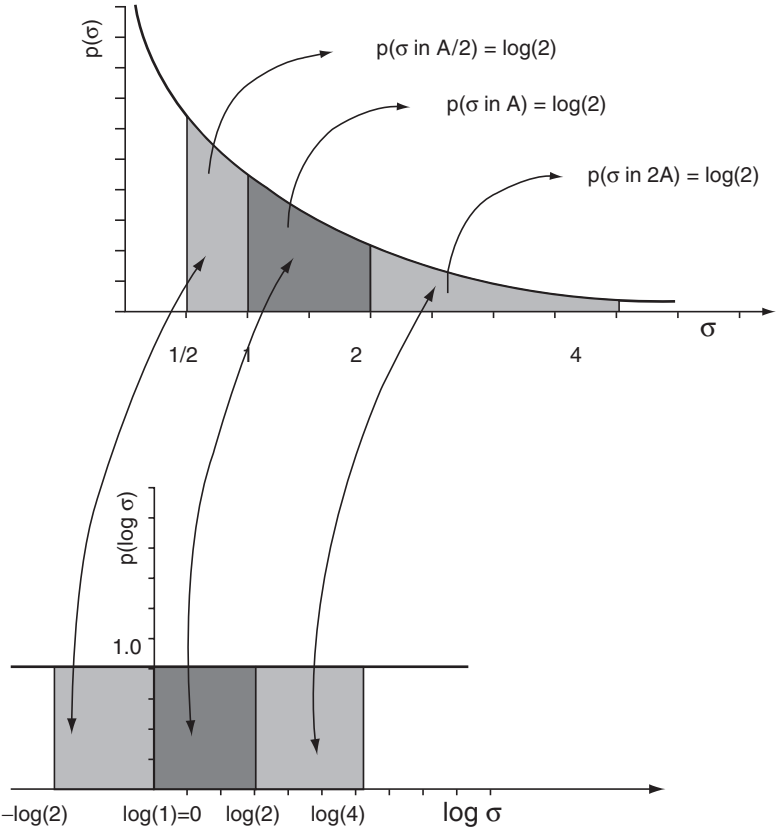


Figure 11.5. Illustration of invariance of noninformative prior for a scale parameter $\sigma > 0$ and the corresponding prior in the transformed log σ space. In the figure, A=(1, 2)

11.7.3 Jeffreys' Non-informative Priors**

The idea of using a formal rule to define a non-informative prior is due to Jeffreys[7]. He used the concept of invariance as a formal rule. Thus far we have not used a well-defined notion of invariance. We now define it more formally.

Invariance principle. If some rule led to $p(\theta)$ as a noninformative prior for θ, the same rule should lead to

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| \tag{11.8}$$

[7]This section contains somewhat more advanced material and may be skipped on a first reading.

as a noninformative prior for ϕ, where $\phi = h(\theta)$ is a one to one transformation. If this is true, posterior inferences based on $p(\phi)$ will be the same as those made on $p(\theta)$.

Thus, what we request is that the conclusions we reach in our analysis do not change if we transform the parameter. This occurs only if the prior for a transformed quantity is consistent with the transformation of random variables formula (see Appendix D). The principle implies, in particular, that regardless of the origin and scale of measurements of quantities of interest, which are always arbitrary, the conclusions that we reach or inferences we make will be invariant.

Jeffreys then showed that a prior that meets the invariance principle is:

$$p(\theta) \propto I(\theta)^{1/2} \tag{11.9}$$

where $I(\theta)$ is Fisher's information for the parameter θ, defined by:

$$I(\theta) = -E\left[\frac{d^2 \log\, p(y|\theta)}{d\theta^2} |\ \theta\right].$$

This criterion, when applied individually to location and scale parameters as in the previous section, results in $p(\theta) \propto$ constant and $p(\sigma) \propto 1/\sigma$, respectively[8].

It is easy to see[9] that Jeffreys criterion (11.9) satisfies the invariance principle (11.8).

Suppose $\phi = h(\theta)$, and write $p = p(y|\theta)$. Then, from the chain rule for differentiation,

$$\frac{d\ \log p}{d\theta} = \frac{d\ \log p}{d\phi}\ \frac{d\phi}{d\theta}.$$

Therefore,

$$I(\theta) = -\frac{d\phi}{d\theta} E\left(\frac{d\ \log p}{d\phi}\ \ \frac{d\ \log p}{d\phi}\right)\frac{d\phi}{d\theta} = I(\phi)\left(\frac{d\phi}{d\theta}\right)^2.$$

Thus,

$$I(\theta)^{1/2} = I(\phi)^{1/2}\left|\frac{d\phi}{d\theta}\right|.$$

[8]It is interesting to note that for the binomial parameter p (proportion), the prior we recommended earlier is U(0,1) (equivalent to a Beta(1,1)) but Jeffreys' criterion (11.9) yields a Beta(1/2,1/2).

[9]This proof is presented in Zellner [162] who attributed it to M. Stone.

This implies that if we set $p(\theta) \propto I(\theta)^{1/2}$ and use the same rule for setting a prior on ϕ, then

$$p(\phi) \propto \ I(\phi)^{1/2} = I(\theta)^{1/2} \left| \frac{d\theta}{d\phi} \right| \propto p(\theta) \left| \frac{d\theta}{d\phi} \right|$$

and this satisfies the invariance principle.

For multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'$ Jeffreys' prior is

$$p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2} \tag{11.10}$$

(the square root of the determinant of Fisher's information matrix), where

$$I_{ij} = -E \left[\frac{\partial \log \, p(y|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right], \ i, j = 1, 2, \ldots, k.$$

However, when $I(\boldsymbol{\theta})$ is not diagonal, this would imply that *a priori*, the parameters are dependent, and this is counter to our intuition of a noninformative prior on the parameters. Therefore, Jeffreys suggests that instead of using criterion (11.10), one should use his scalar criterion (11.9) one parameter at a time. This implies we assume the multiple parameters are *a priori* independent, and individually noninformative, but do not follow jointly the concept of what a "non-informative prior" should be according to Jeffreys himself.

Jeffreys made some efforts to justify the prior $1/\sigma$ for a scale parameter $\sigma > 0$ on grounds other than on invariance. In particular, since

$$\int_0^\infty \frac{d\sigma}{\sigma} = \infty$$

he argued that the value to represent certainty is infinity, rather than unity, which is an equally arbitrary choice. Furthermore, he then notices that

$$\int_0^a \frac{d\sigma}{\sigma} = \int_a^\infty \frac{d\sigma}{\sigma} = \infty$$

which imply that

$$\frac{P(0 < \sigma < a)}{P(a < \sigma < \infty)}$$

is undetermined. This, Jeffreys argued, was a nice indication of our lack of information about σ, since in contrast, for the prior $p(\sigma) \propto$ constant we have that this ratio is zero, an indication that we do know something about σ.

This line of argument is not very solid, since for Jeffreys prior for a scale parameter[10],

$$\frac{P(a \leq \sigma \leq b)}{P(b < \sigma < \infty)} = 0$$

so it is more probable that $\sigma > b$ than $a \leq \sigma \leq b$. Therefore, the best justification of Jeffreys' priors is through the invariance principle.

11.7.4 Other Approaches to Setting Non-informative Priors

As it should be clear from the previous section, one problem with non-informative priors of the type suggested by the invariance principle is that a prior that is flat or uniform in one parametrization (i.e., transformation of parameters) might not be so in another (e.g., consider the case of the prior for a scale parameter). A question is: on which parametrization should we assign a uniform or flat prior? In other words, on which transformed parameter space should $p(h(\theta))$ be proportional to a constant?

Box and Tiao [28] suggested an answer to this question. They noted that the transformation $h(\theta)$ over which we should have a "locally uniform prior"[11] should make the likelihood *data translated*. That is, the likelihood changes with the transformation, and what we require according to this criterion is that only the location of the likelihood changes with the data in the transformed space defined by $h(\theta)$. A data translated likelihood has the form $p(\theta|y) = g(h(\theta) - f(y))$ where g is a known function of the data y. Thus, these authors suggest to pick a $h(\theta)$ that achieves this type of likelihood, and then set $p(h(\theta)) \propto$ constant.

Finding a transformation that yields a data-translated likelihood is not always possible, e.g., in some multiparameter applications. For the Normal distribution discussed above this criterion does yield the same priors as the invariance principle. In multiparameter situations Box and Tiao suggest to rely on some other argument leading to Jeffreys' rule ([28, p. 53]).

Other approaches to defining a non-informative prior have been derived from the idea of measuring the information of the prior distribution using Shannon's entropy. This certainly makes sense since defining what we mean by "lack of

[10] This counterargument was made by J. Neyman [120], who was not fond of Bayesians.

[11] "These authors refer to the improper U$(-\infty, \infty)$ prior as "locally uniform", following Jeffreys.

information" is the fundamental difficulty faced in objective Bayesian analysis. See [74] for this and other approaches for setting priors in Bayesian analysis.

11.8 Inferences on Normally Distributed Data (Known Variance)

We know present some examples of Bayesian inference to illustrate the previous developments. Consider data that is normal: $Y \sim N(\theta, \sigma^2)$ with σ^2 known and θ unknown. This is probably not very realistic in practice, but it is a useful first simple model to illustrate the ideas. Suppose, also for simplicity, that we observe *one* data point. The *likelihood* for one observation is

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{\frac{1}{2\sigma}(y-\theta)^2}.$$

11.8.1 Analysis with a Conjugate Prior

The *conjugate prior* for normal data is

$$\theta \sim N(\mu_0, \tau_0^2)$$

where we assume that the "hyperparameters" μ_0 and τ_0 are known. That this is a conjugate prior for the normal can be seen because the posterior density is an exponential in a quadratic form in θ, which turns out to be a normal:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \propto e^{-\frac{1}{2}\left[\left(\frac{y-\theta}{\sigma}\right)^2+\left(\frac{\theta-\mu_0}{\tau_0}\right)^2\right]}$$

which can be simplified by completing the square[12] in θ to:

$$p(\theta|y) \propto e^{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2}, \quad \text{thus } \theta|y \sim N(\mu_1, \tau_1^2)$$

where the posterior mean is:

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}}$$

and the posterior variance follows the relation:

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}.$$

[12]Recall this means we get a perfect binomial square of the form $(\theta - c)^2$. Once this is done, we treat anything not a function of the random variable as a constant, which is not shown due to the use of the proportionality sign.

Thus the posterior mean is a weighted average of the prior mean and data, with weights equal to the precisions (i.e., the inverse of the variances). The posterior precision is the sum of the prior precision and data precision. The posterior mean can also be interpreted as:

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

so we can say that the prior mean is adjusted towards the observed y. Similarly,

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

so some authors (e.g., [60]) say that the data "shrinks" towards the prior mean. Some interesting cases are:

- If $\tau_0^2 = 0$ then $\mu_1 = \mu_0$, i.e., the prior mean is "infinitely precise" and dominates;
- if $\sigma^2 = 0$ then $\mu_1 = y$, i.e., the data is "infinitely precise" and dominates;
- if $y = \mu_0$ then $\mu_1 = \mu_0 = y$, i.e., the data and prior means agree and so does the posterior mean;
- if $\tau_0^2 \to \infty$ then $\mu_1 \to y$, i.e., we approach a "non-informative" prior on the mean parameter.

The *posterior predictive density* is obtained from

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \propto \int e^{-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2}\, e^{\frac{1}{2\tau_1^2}(\theta-\mu_1)^2} d\theta \\ &\propto \int e^{\frac{1}{2}\theta^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_1^2}\right)+\theta\left(\frac{\tilde{y}}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)-\frac{1}{2}\left(\frac{\tilde{y}^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)} d\theta \end{aligned}$$

Integrating with respect to θ yields

$$p(\tilde{y}|y) \propto e^{-\frac{1}{2}\left(\frac{\tilde{y}^2}{\sigma^2}+\frac{\mu_1^2}{\tau_1^2}\right)}\, e^{\frac{1}{2}\left(\frac{\tilde{y}}{\sigma^2}+\frac{\mu_1}{\tau_1^2}\right)^2\left(\frac{1}{1/\sigma^2+1/\tau_1^2}\right)}$$

Completing the square in $\tilde{y}$ in the exponent gives:

$$p(\tilde{y}|y) \propto e^{-\frac{1}{2}\frac{(\tilde{y}-\mu_1)^2}{\sigma^2+\tau_1^2}}$$

Therefore, the posterior predictive density is given by

$$\tilde{y}|y \sim N(\mu_1, \sigma^2 + \tau_1^2).$$

Case of several observations. If n independent and identically normally-distributed data points are observed, the *posterior* of the mean is

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) = p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\ &\propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2} \prod_{i=1}^{n} e^{-\frac{1}{2\sigma^2}(y_i-\theta)^2} \\ &\propto e^{-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]} \end{aligned}$$

Completing the square on θ (placing terms not a function of θ in the proportionality constant):

$$P(\theta|y_1, y_2, \ldots, y_n) = p(\theta|\overline{y}) = N(\mu_n, \tau_n^2)$$

where the first equality follows because the sample mean $\overline{y}$ is a sufficient statistic, i.e., the posterior is only a function of the data through the sample average. Here we have that

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

and the posterior precision is:

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}.$$

Note how as $n \to \infty$ or as $\tau_0 \to \infty$, $\mu_n \to \overline{y}$ and $\tau_n^2 \to \sigma^2/n$. This coincides with the frequentist results.

Example. Inferences on normal distributed data, variance known. Suppose $\sigma^2 = 50$ is known from previous experience, and we take $n = 5$ observations from which $\overline{y} = 370$. Figure 11.6 shows the results for a conjugate prior $\theta \sim N(500, 2^2)$ distribution. The prior is very precise, and the posterior has a mean close to the prior mean ($\mu_n = 462.85$, $\tau_n = 1.69$). The posterior predictive density also has mean equal to $\mu_n = 462.85$ but its

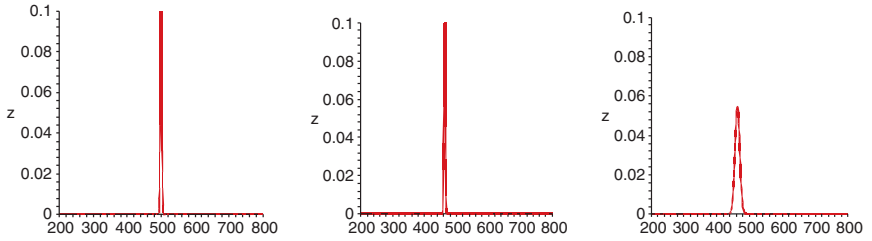


Figure 11.6. Inferences on normally distributed data, σ^2 known, "informative" conjugate prior, $\overline{y} = 370$, $n = 5$. Left: $p(\theta)$; center: $p(\theta|y)$; right: $p(\tilde{y}|y)$

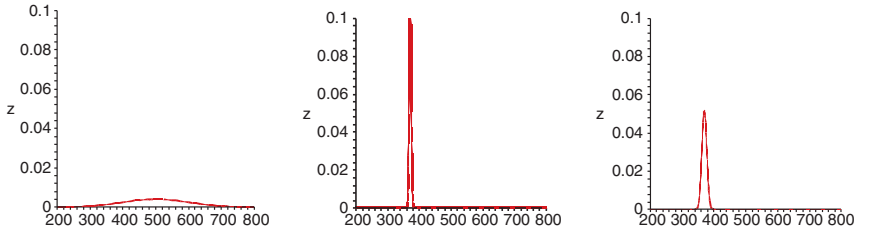


Figure 11.7. Inferences on normally distributed data, σ^2 known, "less informative" conjugate prior, $\overline{y} = 370$, $n = 5$. Left: $p(\theta)$; center: $p(\theta|y)$; right: $p(\tilde{y}|y)$

standard deviation is $\sqrt{\sigma^2 + \tau_1^2} = 7.27$. Therefore, the prior dominates the likelihood.

In contrast, Figure 11.7 shows the corresponding results for a conjugate prior $\theta \sim N(500, 100^2)$, a much flatter distribution. The prior is relatively imprecise, and the posterior has a mean close to the data mean but with larger variability ($\mu_n = 370.12$, $\tau_n = 3.16$). The posterior predictive density also has mean equal to $\mu_n = 370.12$ with standard deviation equal to $\sqrt{\sigma^2 + \tau_1^2} = 7.74$. Here, the data dominates the likelihood. ■

11.8.2 Analysis with a Non-informative Prior

For the normal model with σ^2 known, consider now the non-informative *prior*

$$p(\theta) \propto \text{ constant.}$$

The *posterior* after n observations is

$$\begin{aligned} p(\theta|y_1, y_2, \ldots, y_n) &\propto p(y_1, y_2, \ldots, y_n)p(\theta) \\ &= p(y|\theta) \end{aligned}$$

$$\begin{aligned} &\propto \quad e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2} \\ &= \quad e^{-\frac{1}{2\sigma^2}[n\theta^2-2\theta\sum y_i+\sum y_i^2]} \\ &= \quad e^{-\frac{n}{2\sigma^2}(\theta^2-2\theta\overline{y}+\sum_{i=1}^{n} y_i^2/n)}. \end{aligned}$$

Completing the square on θ in the exponent (placing constant terms in the proportionality constant):

$$p(y|\theta) \propto e^{-\frac{n}{2\sigma^2}(\theta-\overline{y})^2}.$$

Therefore,

$$\theta|y \sim N(\overline{y}, \sigma^2/n) \tag{11.11}$$

which evidently is a proper posterior. Note how we also obtain this distribution in the conjugate prior case when $\tau_0 \to \infty$.

11.9 Inferences on Normally Distributed Data, Both Parameters Unknown

We will only sketch the main results for a non-informative prior. For two unknown parameters, we apply Jeffreys rule one parameter at-a-time, i.e., we assume *a priori* that the parameters are independent:

$$p(\theta) \propto \text{ constant} \quad \text{and} \quad p(\sigma^2) \propto 1/\sigma^2$$

which yields

$$p(\theta, \sigma^2) = p(\theta)p(\sigma^2) \propto 1/\sigma^2.$$

The *joint posterior distribution* is obtained by making use of the expression:

$$p(\theta, \sigma^2|y) = p(\theta|\sigma^2, y)p(\sigma^2|y)$$

where, if σ^2 is given, we have shown (see equation (11.11)) that for a noninformative prior on θ

$$\theta|\sigma^2, y \sim N(\overline{y}, \sigma^2/n).$$

We can also show[13] that

$$\sigma^2|y \sim Inv-\chi^2(n-1, s^2).$$

[13]The Inv$-\chi^2(v_0, \sigma_0^2)$ (scaled inverse chi-squared) distribution is the distribution of $\sigma_0^2 v_0^2/\chi_{v_0}^2$, i.e., it is the inverse of a usual χ^2 distribution with v_0 degrees of freedom that is scaled by the quantity $\sigma_0^2 v_0^2$, hence its name. This is a particular case of an inverse gamma distribution.

With these two distributions, the joint posterior can easily be simulated. (Note how the distribution of $\sigma^2|y$ is analogous to the classical (frequentist) result $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ where s^2 is the random variable. In the Bayesian case, σ^2 is the random variable.)

The marginal posterior for the mean is

$$\theta|y \sim t_{n-1}(\overline{y}, s^2/n)$$

where the statistics $\overline{y}$ and s^2 are sufficient. The distribution is a noncentral t distribution with location parameter $\overline{y}$ and scale parameter s^2/n. (Note this implies the Bayesian result:

$$\frac{\theta - \overline{y}}{s/\sqrt{n}} \sim t_{n-1}$$

a central t, where θ is the random variable. Compare this with the classical result:

$$\frac{\overline{y} - \theta}{s/\sqrt{n}} \sim t_{n-1}$$

where $\overline{y}$ is the random variable.)

The *posterior predictive distribution* is given by:

$$p(\tilde{y}|y) = \int\int \underbrace{p(\tilde{y}|\theta, \sigma^2, y)}_{N(\theta,\sigma^2)} p(\theta, \sigma^2|y) d\theta d\sigma^2$$

which can be approximated by simulation:

1 Draw (θ, σ^2) from $p(\theta, \sigma^2|y)$:

   (a) draw σ^2 from $\sigma^2|y \sim Inv - \chi^2(n-1, s^2)$;

   (b) draw θ from $\theta|\sigma^2, y \sim N(\overline{y}, \sigma^2/n)$;

2 draw $\tilde{y}$ from $\tilde{y} \sim N(\theta, \sigma^2)$

3 goto 1 until we have iterated N times.

The integral can also be solved analytically, after some algebra, yielding

$$\tilde{y}|y \sim t_{n-1}\left(\overline{y}, \left(1+\frac{1}{n}\right)s^2\right).$$

If known in closed-form, the predicted density can be simply graphed and reported. Alternatively, it can be simulated as the following example shows.

Example, inferences on normal data, both parameters unknown. Suppose we collect 10 observations from which $\overline{y} = 100$ and $s^2 = 20$. We then simulate the predictive density of a new observation, $\tilde{y}|y$, using a) the distributions of $\sigma^2|y$ and $\theta|\sigma^2, y$, and b) using the t distribution directly. Table 11.3 shows the corresponding Matlab code. Figure 11.8 shows histograms obtained using the marginal distributions; Figure 11.9 shows the corresponding posterior predictive density generated directly from the t distribution's closed form. The two simulated predictive densities are practically the same as expected. The distribution of $\sigma^2|y$ is an scaled inverse chi-square, with expected value $v_0/(v_0-2)s^2 = 25.7$ and mode at $v_0/(v_0+2)s^2 = 16.36$ where $v_0 = n-1 = 9$. Although in this case using the closed form of the predictive

Table 11.3. Matlab simulation program for the posterior predictive density $\tilde{y}|y$, normal data, both parameters unknown, non-informative prior. The program uses both the distributions of $\sigma^2|y$ and $\theta|\sigma^2, y$ and the t distribution directly

```
function [ynew1,ynew2,theta,sigma2]
     =predictiveDensity(noOfSimulations,ybar,n,s2);
for i=1:1:noOfSimulations
    x=chi2rnd(n-1);
    sigma2(i)=(n-1)*s2/x;
    theta(i)=normrnd(ybar,sqrt(sigma2(i)/n));
    ynew1(i)=normrnd(theta(i),sqrt(sigma2(i)));
    % Use instead the direct formula (a non-central
        t dist.)
    ynew2(i)=ybar+sqrt(s2*(1+1/n))
      *normrnd(0,1)*sqrt(n)/sqrt(x);
end;
```

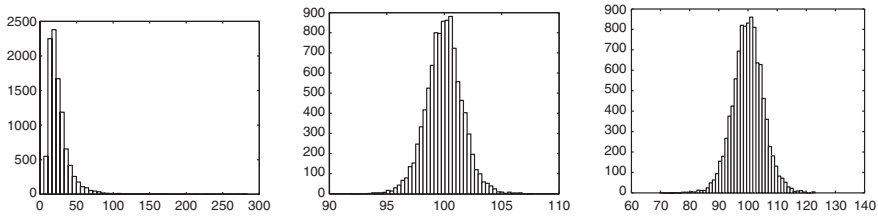


Figure 11.8. Inferences on normally distributed data, θ and σ^2 unknown, non-informative prior, $\overline{y} = 100$, $s^2 = 20$, $n = 10$. Left: $p(\sigma^2|y)$; center: $p(\theta|\sigma^2, y)$; right: $p(\tilde{y}|y)$

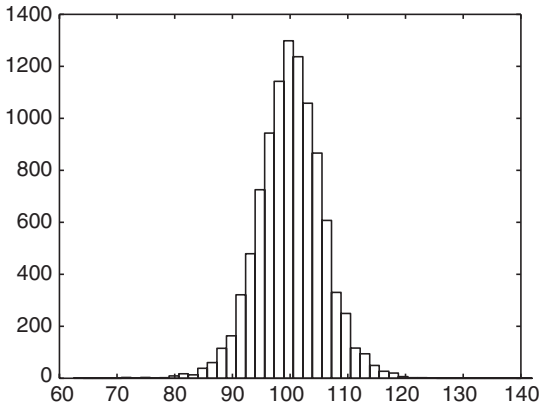


Figure 11.9. Predictive density ($p(\tilde{y}|y)$) obtained by simulating directly from the t distribution, θ and σ^2 unknown, non-informative prior, $\overline{y} = 100$, $s^2 = 20$, $n = 10$. Compare with Figure 11.8 c)

density is easy, this illustrates a useful approach to generate the distribution of $\tilde{y}|y$ when no closed-form expression exists. ■

11.10 Problems

1 Consider the first Matlab program for the determination of the posterior of the CV of a normal with both parameters unknown (we used non-informative priors). Modify it such that you can compute the posterior distribution of the C_{pk} statistic, a common "capability" statistic used in quality control defined by:

$$C_{pk} = \frac{\min\{(USL - \theta), (\theta - LSL)\}}{3\sigma}$$

where USL and LSL are two known constants (lower and upper specification limits of the quality characteristic y). The response or quality characteristic y is assumed $N(\theta, \sigma^2)$ with both parameters unknown. Assume non-informative priors in both parameters. Give a listing of the program and sample output for the following case: $LSL = 38, USL = 62$, and a sample of size 15 yield $\overline{y} = 53.1$ and $s^2 = 1.95$. Use a histogram to show the simulated distribution. [Note: this shows the power of the Bayesian approach, as the sampling (frequentist) distribution of this statistic is rather complex].

2 An engineer uses a measuring gauge with a known standard deviation of 0.12 to take nine independent measurements of a quality characteristic. The measurements can be reasonably assumed $N(\theta, 0.12^2)$ with unknown mean. The sample mean of the observations is 17.653. Find a 99% credible interval on the (future) tenth measurement (this is called a Bayesian tolerance interval). Assume a non-informative prior.

3 Suppose $Y \sim N(\theta, 1)$ and we take three independent observations $y_1 = 2$, $y_2 = 3$, and $y_4 = 4$. Assuming your prior for the mean θ is $N(3.6, 2^2)$, find the posterior distribution of θ.

4 Suppose we measure the tensile strength of a metal specimen for which it is known its mean tensile strength is 50 units. We are interested in determining the variance of the tensile strength measurements for a capability analysis. The measurements are assumed $N(50, \sigma^2)$. Twenty tensile strength measurements were taken from which $v = \sum_{i=1}^{20}(y_i - 50)^2/20 = 17.4$. Using a non-informative prior, find the mode of the posterior distribution of σ^2.

Chapter 12

BAYESIAN METHODS FOR PROCESS OPTIMIZATION

What experimenters take for granted before they begin their experiments is infinitely more interesting than any result to which their experiments lead.

—Norbert Weiner (1894–1964)

The mainstream literature on Response Surface Optimization is classical or "frequentist" given that it considers parameters as unknown constants that need to be estimated from data. The sampling variability or experimental error is reflected in the sampling distributions of the estimates. This sampling variability can (and should) be considered in optimization, see Chapter 7. In contrast, the Bayesian approach to statistical inference considers model parameters (and in fact, any unknowns) as random variables. This has considerable advantages over the classical approach when optimizing a process based on a fitted model, since depending on the estimated parameters different optimal conditions will be determined. In the Bayesian approach, the uncertainty in the model's parameters is directly incorporated in the analysis. Prior knowledge (which can be considerable, in agreement with Weiner's quote above) can be incorporated, if desired, into the optimization process. Otherwise, non-informative priors can be used for optimization purposes.

This chapter presents Bayesian linear regression models and its use in process optimization, with the minimum number of technical details, without sacrificing understanding of the main ideas for readers not familiar with Bayesian Analysis.

In this chapter, we will first give results needed to understand Bayesian regression models and their use in optimization problems. One such problem is:

$$\max_{\boldsymbol{w}\in R}\int_A p(\tilde{\boldsymbol{y}}|data,\boldsymbol{w})d\tilde{\boldsymbol{y}}$$

where $\tilde{\boldsymbol{y}}$ is a $q\times 1$ vector of future responses, $\boldsymbol{w}$ is a vector of controllable factors, and A is a specification region for the q responses. Solving such optimization problem provides a solution that satisfies the specifications or tolerances on the responses of interest. We consider application of these ideas to factorial and mixture experiments. A related problem is when to stop experimenting, which can be solved using a Bayesian approach due to Gilmour and Mead [61], and is presented here.

Another problem we will discuss from a Bayesian point of view is that of **Robust Parameter Design**. Here, noise factors $\boldsymbol{z}$ are treated as random variables, so optimization of the problem above is carried out with an additional integration over the noise factors, i.e., $\boldsymbol{w}$ is split into $(\boldsymbol{w}_c,\boldsymbol{w}_{nc})$, the controllable (optimizable) factors and the non-controllable factors; the latter are integrated out of the objective function according to some prior distribution.

Finally, we would like to investigate Bayesian *model-robust optimization* problems. Suppose we have a family of m potential models M_i that adequately explain some process. Optimizing each model independently may lead to very different solutions. If we can compute $P(M_i|data)$, the posterior of model i, then we could solve[1]

$$\max_{\boldsymbol{w}\in R}\sum_{\text{all } i}\int_A p(\tilde{y}|M_i,data,\boldsymbol{w})d\tilde{y}\; p(M_i|data),$$

the model-averaged posterior probability of satisfying the constraints A. We will focus on the case of a single response of interest and on models linear in the parameters.

12.1 Bayesian Linear Regression and Process Optimization

Consider a normal linear regression model with p regressors $x_1,x_2,\ldots,x_p$ which themselves can be nonlinear transformations of underlying controllable factors. For example, $x_3=x_1x_2$, or $x_4=\log(x_2)$, etc. We assume we have

[1]As we will see, the solution to this problem is not equivalent to averaging the solutions $\boldsymbol{w}_i^*$ obtained from optimizing each model M_i independently.

conducted N "experiments", an experiment consisting on observing the values of the p regressors together with the value of the response, y. In this section, we assume a single response is of interest. The N experimental conditions are gathered in a $N \times p$ matrix

$$\boldsymbol{X} = [x_{ij}]$$

which we will call the "design matrix" in what follows (this will include the actual $N \times k$ design matrix if there are k underlying factors). Put the observed response values in an $N \times 1$ vector $\boldsymbol{y}$. If the assumed normal model is valid, it should be valid for all N observations, so for each observation i we can express the response as

$$y_i|\boldsymbol{\beta}, X = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where we assume $\varepsilon \sim N(0, \sigma^2)$ and $\boldsymbol{\beta}$ denotes the $p \times 1$ vector of parameters. The response is then assumed to be the result of two effects: the first one, which we can explain, is due to the p regressors, the second one, which we cannot attribute to any of the p factors, we thus model it as a random variable with mean zero and constant variance. The two sources of uncertainty, due to not knowing the parameters, and due to not knowing the intrinsic variability of the errors of the model, need to be considered in any inference problem. This is achieved by using the Bayesian approach.

Based on this model we have that

$$E[y_i|\boldsymbol{\beta}, \boldsymbol{X}) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

and

$$\text{Var}(y_i\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) = \sigma^2.$$

In many applications, x_1 is assumed to be one, and the model has an intercept[2]. We can summarize the model by saying that

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

12.1.1 Analysis with a Non-informative Prior

Under a standard (Jeffreys) non-informative prior, Bayesian estimates and standard errors coincide with classical results. However, even this non-informative case is of considerable value for *predictive* inference, in particular, for applications in process optimization, as will be shown later.

[2]Models for Mixture experiments usually have no intercept.

The *non-informative prior* for the parameters of this regression model is a Uniform distribution in $(\boldsymbol{\beta}, \log \sigma)$ or

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}) \propto \frac{1}{\sigma^2}.$$

We assume both parameters are independent *a priori*. Note that we indicate that the design is given, since we can design the experiment before deciding on the prior. Because of this, in regression cases it is also customary to define the prior taking advantage of our knowledge of the design. The resulting design is called a "g-prior", and was proposed by Zellner [161]. We will use g-priors in our discussion of model-robust optimization. In this section we instead use non-informative priors.

The likelihood for the N observations under the assumed Normal model is

$$p(\boldsymbol{y} | \boldsymbol{\beta}, \sigma^2, \boldsymbol{X}) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}$$

so the posterior ($\propto$ prior $\times$ likelihood) is

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) \propto \frac{1}{(\sigma^2)^{N/2+1}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}. \tag{12.1}$$

The marginals for each parameter can be obtained by integration. For example, the marginal density for $\boldsymbol{\beta}$ is obtained from

$$\begin{aligned} p(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{X}) &\propto \int_0^\infty \frac{1}{(\sigma^2)^{N/2+1}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})} d\sigma^2 \\ &\propto [(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})]^{-N/2} \end{aligned}$$

Add and subtract $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ in each parenthesis, where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$ makes the right hand side equal to

$$\begin{aligned} &= [(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})]^{-N/2} \\ &= \left[(N-p)s^2 + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right]^{-N/2} \\ &= \left[1 + \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{(N-p)s^2}\right]^{-N/2} \underbrace{[(N-p)s^2]^{N/2}}_{\text{a constant}} \\ &\propto \left[1 + \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{(N-p)s^2}\right]^{-(N-p+p)/2} \underbrace{(s^2|\boldsymbol{X}'\boldsymbol{X}|^{-1})^{-1/2}}_{\text{another constant}} \end{aligned}$$

which has the form of a multivariate t distribution (see Appendix D) with degrees of freedom $v = N - p$:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim t_{N-p}(\widehat{\boldsymbol{\beta}}, s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Similarly, we can obtain the posterior marginal distribution of σ^2 by integrating (12.1) over $\boldsymbol{\beta}$, obtaining a scaled inverse chi-square distribution (see Appendix D):

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim Inv\ \chi^2(n-p, s^2).$$

The posterior predictive density is obtained as follows[3]. Let $\tilde{y}$ be a new observation (a scalar) to be obtained at point $\boldsymbol{w}$, a $p \times 1$ vector. We then need to compute:

$$\begin{aligned} & p(\tilde{y}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) \\ & \quad = \int \int_0^\infty p(\tilde{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) d\sigma^2 d\boldsymbol{\beta} \\ & \quad \propto \int \int_0^\infty \frac{1}{(\sigma^2)^{(N+1)/2+1}}\ e^{-\frac{1}{2\sigma^2}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+(\tilde{y}-\boldsymbol{w}'\boldsymbol{\beta})^2]} d\sigma^2 d\boldsymbol{\beta} \\ & \quad \propto \int \frac{d\boldsymbol{\beta}}{[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + (\tilde{y}-\boldsymbol{w}'\boldsymbol{\beta})^2]^{(N+1)/2}} \end{aligned}$$

where

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})' = (N-p)s^2 + (\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})$$

and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

A common trick in Bayesian statistics is to identify the integrand of an integral that looks hard as a known density. If this is possible, it provides a mechanism to find a closed-form expression for the integral, as we now show by example.

The denominator in the last integral can be written as

$$[\boldsymbol{C}_3 + (\boldsymbol{\beta}-\boldsymbol{C}_1)'\boldsymbol{C}_2(\boldsymbol{\beta}-\boldsymbol{C}_1)]^{(N+1)/2}$$

[3]We follow the derivation by Press [129] to do this.

where

$$\begin{aligned}
\underbrace{\boldsymbol{C}_1}_{p\times 1} &= (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{w}\boldsymbol{w}')^{-1}(\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \tilde{y}\boldsymbol{w}) \\
\underbrace{\boldsymbol{C}_2}_{p\times p} &= \boldsymbol{X}'\boldsymbol{X} + \boldsymbol{w}\boldsymbol{w}' \\
\underbrace{\boldsymbol{C}_3}_{1\times 1} &= (N-p)s^2 + \frac{(\tilde{y} - \boldsymbol{w}'\widehat{\boldsymbol{\beta}})^2}{1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}.
\end{aligned}$$

Thus we can write

$$p(\tilde{y}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) \propto \int \frac{d\boldsymbol{\beta}}{[\boldsymbol{C}_3 + (\boldsymbol{\beta} - \boldsymbol{C}_1)'\boldsymbol{C}_2(\boldsymbol{\beta} - \boldsymbol{C}_1)]^{(N+1)/2}}$$

which has an integrand with the form of the Kernel [4] of a matrix T distribution (see Appendix D). Since the multiple integral is over all possible values of $\boldsymbol{\beta}$, and the integral of the kernel times the constant of a matrix T equals to one, the integral of the kernel (only, without the constant) must equal the inverse of the constant of a multivariate T, namely (using the notation from Appendix D):

$$\frac{1}{\frac{k(m,l_1,l_2)}{|\boldsymbol{P}|^{(m-l_2)/2}|\boldsymbol{Q}|^{l_2/2}}}$$

where we have the equivalences $\boldsymbol{P}^{-1} \Leftrightarrow \boldsymbol{C}_3$, $T \Leftrightarrow (\boldsymbol{\beta} - \boldsymbol{C}_1)'$, $\boldsymbol{Q}^{-1} \Leftrightarrow \boldsymbol{C}_2$, $l_1 = 1$, $l_2 = p$, and $m = N + 1$. Using these equivalences in the latter expression, we notice that only $\boldsymbol{C}_3^{-1}$ (i.e., $\boldsymbol{P}$) is a function of $\tilde{y}$, the random variable, so everything else can be treated as a constant. We thus have that

$$p(\tilde{y}|\boldsymbol{y}) \propto \frac{1}{|\boldsymbol{C}_3|^{(N+1-p)/2}} = \frac{1}{\boldsymbol{C}_3^{(N+1-p)/2}}$$

where recall that

$$\boldsymbol{C}_3 = (N-p)s^2 + \frac{(\tilde{y} - \boldsymbol{w}'\widehat{\boldsymbol{\beta}})^2}{1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}.$$

Therefore, the predictive density is

$$p(\tilde{y}|\boldsymbol{y}) \propto \frac{1}{\left[1 + \frac{1}{N-p}\frac{(\tilde{y} - \boldsymbol{w}'\widehat{\boldsymbol{\beta}})^2}{s^2\left(1 + \boldsymbol{w}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{w}\right)}\right]^{(N+1-p)/2}}$$

[4] The Kernel of the density is the part of the density function which is a function of the random variable, the rest being a constant or functions or constants.

which is a univariate Student t density (see Appendix D):

$$\tilde{y}|\boldsymbol{y} \sim t_{N-p}(\boldsymbol{w}'\widehat{\boldsymbol{\beta}}, s^2(1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w})).$$

Note that this implies that

$$\text{Var}(\tilde{y}|\boldsymbol{y}) = \frac{N-p}{N-p-2}s^2(1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w})$$

which is the sum of two components: the first one, proportional to s^2, is due to intrinsic (sampling) variability; the second one, proportional to $s^2(\boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w})$, represents variance due to the uncertainty in the parameters.

If we want to predict the next several M observations, it can be shown, following a similar approach, that the vector of predicted responses $\tilde{y}$ has density

$$\tilde{\boldsymbol{y}}|\boldsymbol{y} \sim t_{N-p}(\boldsymbol{W}\widehat{\boldsymbol{\beta}}, s^2(\boldsymbol{I} + \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{W}'))$$

which is an M-dimensional student t density where $\tilde{\boldsymbol{y}}$ is a $M \times 1$ vector of future values of the response and $\boldsymbol{W}$ is a $M \times p$ matrix that gives the M points at which we wish to predict, where the columns of $\boldsymbol{W}$ are in model form in correspondence with the columns of $\boldsymbol{X}$.

A final useful result is the *predictive region* for a vector of future responses $\tilde{\boldsymbol{y}}$, given by:

$$\frac{(\tilde{\boldsymbol{y}} - \boldsymbol{W}\widehat{\boldsymbol{\beta}})'(\boldsymbol{I}_M + \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{W}')^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{W}\widehat{\boldsymbol{\beta}})}{(N-p)s^2} \sim F_{M,N-p}.$$

Points $\tilde{\boldsymbol{y}}$ of the space of the responses that have a left hand expression less than $F_{M,N-p,\alpha}$ constitute the $(1-\alpha)100\%$ predictive region. The interpretation of this region is the Bayesian one: this region contains the predicted response values with probability $1-\alpha$.

Example. Computation of a predictive region. Montgomery [110] gives a chemical experiment in which three responses are of interest: the yield of the process, the viscosity of the chemical, and the molecular weight. In this example we will consider the yield response. Two underlying factors are controllable, the reaction time and the temperature, and we fit a quadratic polynomial. The $\boldsymbol{X}$ matrix containing the design (a rotatable CCD) and the observed responses are shown in Table 12.1.

Table 12.1. Data for a chemical experiment

Int.	x_1	x_2	x_1x_2	x_1^2	x_2^2	y_1=yield	y_2=viscosity
1	−1	−1	1	1	1	76.5	62
1	−1	1	−1	1	1	77	60
1	1	−1	−1	1	1	78	66
1	1	1	1	1	1	79.5	59
1	0	0	0	0	0	79.9	72
1	0	0	0	0	0	80.3	69
1	0	0	0	0	0	80	68
1	0	0	0	0	0	79.7	70
1	0	0	0	0	0	79.8	71
1	1.41421	0	0	2	0	78.4	68
1	−1.41421	0	0	2	0	75.6	71
1	0	1.41421	0	0	2	78.5	58
1	0	−1.41421	0	0	2	77	57

Suppose we wish to compute a 95% prediction region for the yield response obtained at the points $(0.389, 00.306)$ (the optimal point estimate obtained from a least squares fit) and $(-1, -1)$, in coded units. The $\boldsymbol{W}$ matrix is then given by

$$\boldsymbol{W} = \begin{bmatrix} 1 & .389 & .306 & .1190 & .1513 & 0.0936 \\ 1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

The 95% prediction region is shown in Figure 12.1. The Maple 9 command:

```
> plot3d(Transpose(<y1,y2>-W.B).MatrixInverse(eye2+W.
MatrixInverse(Transpose(X).X).Transpose(W)).(<y1,y2>-W.B)/
((13-6)*s^2)-8.64,y1=70..90,y2=70..90,view=-100..0);
```

creates the plot, where the user should previously specify the values of `s, B, X, y, W` and the "eye2" matrix (identity matrix of dimension 2). The prediction at the first point evidently results in higher yields than at the second one. The prediction is also more precise at the first point than at the second, as it can be seen for the slightly larger elongation in the vertical direction. However, the region is not too large, a consequence of the excellent fit of the quadratic polynomial. ■

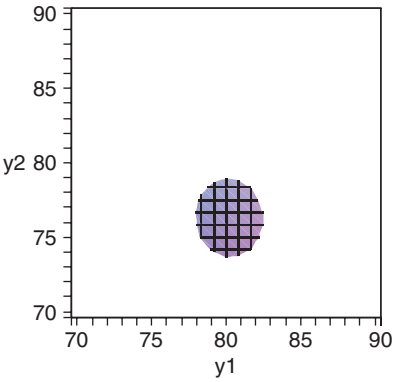


Figure 12.1. 95% prediction region for the yield response at points $(0.389, 0.306)$ and $(-1, -1)$, respectively

Example. Optimization of a single response process. Consider now the use of the predictive density for optimization purposes. Consider again the chemical experiment data of the previous example, and let the response of interest be the viscosity. The process engineer wishes to know the operating conditions that would maximize the probability of the viscosity being between 62 and 68 units. To do this, we need to solve:

$$\max_{\sqrt{2} \le x_1, x_2 \le \sqrt{2}} \int_{62}^{68} p(\tilde{y}|\boldsymbol{y}) d\tilde{y}$$

where the region over we wish to vary the two controllable factors ranges from $-\sqrt{2}$ to $\sqrt{2}$ in coded units, given the central composite design that was used. Section 12.10 discusses the MATLAB (v. 7) program `maxArea.m` that performs the optimization. Since the objective function is not concave, the `fmincon` nonlinear optimization solver is run from a set of initial points selected according to a random latin hypercube. For the viscosity response, we have that the highest probability found is 0.7305 at $w_1 = 0.0852,\ w_2 = 0.7845$, so the solution is well inside the experimental region. ■

12.1.2 Analysis with a Conjugate Prior

The non-informative regression model has the limitation that $N > p$, otherwise, the $\boldsymbol{X}'\boldsymbol{X}$ matrix is not invertible. The conjugate case does not have this limitation, as it incorporates prior information that, when added to $\boldsymbol{X}'\boldsymbol{X}$, makes the resulting matrix invertible. Here we present the derivation of the posterior

distribution of the parameters (including σ^2 as an unknown) using a conjugate prior[5].

The prior distribution can be written as

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$$

where

$$p(\boldsymbol{\beta}|\sigma^2) \propto \frac{\boldsymbol{A}_0^{1/2}}{(\sigma^2)^{p/2}}\ e^{-\frac{1}{2\sigma}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}$$

(or $\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2\boldsymbol{A}_0^{-1})$), and

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{v_0/2+1}}\ e^{-\frac{v_0\sigma_0^2}{2\sigma^2}}$$

(or $\sigma^2 \sim Inv\ \chi^2(v_0, \sigma_0^2)$).

Therefore, the joint prior is a Normal-Inv $\chi^2(\boldsymbol{\beta}_0, \sigma_0^2\boldsymbol{A}_0^{-1}; v_0, \sigma_0^2)$:

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{(v_0+p)/2+1}}\ e^{-\frac{1}{2\sigma^2}[v_0\sigma_0^2+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)]}.$$

The likelihood of n observations is

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}}\ e^{-\frac{1}{2\sigma^2}[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]}.$$

Therefore, the posterior distribution is

$$\begin{aligned} &p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \\ &\quad\propto \frac{1}{(\sigma^2)^{(n+p+v_0)/2+1}}\ e^{-\frac{1}{2\sigma^2}[v_0\sigma_0^2+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})]}. \end{aligned} \tag{12.2}$$

In analogy with the way we parameterized the prior, we should be able to write the posterior as a Normal-Inv $\chi^2(\boldsymbol{\beta}_n, \sigma_n^2\boldsymbol{A}_n^{-1}; v_n, \sigma_n^2)$. To achieve this we develop the products in the exponent:

$$\begin{aligned} &(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{A}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &\quad= \boldsymbol{\beta}'\boldsymbol{A}_0\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{A}_0\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta} + \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta}_0 \\ &\qquad+\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\ &\quad= \boldsymbol{\beta}'(\boldsymbol{A}_0+\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\boldsymbol{A}_0\boldsymbol{\beta}_0+\boldsymbol{X}'\boldsymbol{y}) + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta}_0. \end{aligned} \tag{12.3}$$

[5]We follow Zellner [162] in this section.

Define

$$\boldsymbol{A}_n = \boldsymbol{A}_0 + \boldsymbol{X}'\boldsymbol{X}$$

and complete the square on $\boldsymbol{\beta}$ in (12.3):

$$\begin{aligned}(12.3) &= [\boldsymbol{\beta} - \boldsymbol{A}_n^{-1}(\boldsymbol{A}_0\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y})]'\boldsymbol{A}_n[\boldsymbol{\beta} - \boldsymbol{A}_n^{-1}(\boldsymbol{A}_0\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y})] \\ &\quad + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta}_0 - [\boldsymbol{A}_n^{-1}(\boldsymbol{A}_0\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y})]'\boldsymbol{A}_n\boldsymbol{A}_n^{-1}(\boldsymbol{A}_0\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{\beta} - \boldsymbol{\beta}_n)'\boldsymbol{A}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_n) + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n\boldsymbol{A}_n\boldsymbol{\beta}_n\end{aligned}$$

where

$$\boldsymbol{\beta}_n = \boldsymbol{A}_n^{-1}(\boldsymbol{A}_0\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y}).$$

Therefore, with this exponent and in analogy with the joint prior (12.1), we can write the posterior as

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) \propto \frac{1}{(\sigma^2)^{(n+v_0+p)/2+1}}\, e^{-\frac{1}{2\sigma^2}[v_n\sigma_n^2 + (\boldsymbol{\beta}-\boldsymbol{\beta}_n)'\boldsymbol{A}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_n)]}$$

where

$$v_n\sigma_n^2 = v_0\sigma_0^2 + \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}_0'\boldsymbol{A}_0\boldsymbol{\beta}_0 - \boldsymbol{\beta}_n'\boldsymbol{A}_n\boldsymbol{\beta}_n$$

and

$$v_n = n + v_0.$$

Integrating with respect to σ^2 from 0 to ∞, we get:

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &\propto [v_n\sigma_n^2 + (\boldsymbol{\beta} - \boldsymbol{\beta}_n)'\boldsymbol{A}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_n)]^{-(n+v_0+p)/2} \\ &\propto \left[1 + \frac{1}{v_n}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)'\frac{\boldsymbol{A}_n}{\sigma_n^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_n)\right]^{-(v_n+p)/2} \qquad (12.4)\end{aligned}$$

which has the form of a Student t distribution:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim t_{v_n}(\boldsymbol{\beta}_n, \sigma_n^2\boldsymbol{A}_n^{-1}).$$

Note how the non-informative posterior of $\boldsymbol{\beta}$ is obtained when $v_0 = -p$ and $\boldsymbol{A}_0 = \boldsymbol{\beta}_0 = \sigma_0^2 = 0$. Also, for increasingly larger values of v_0 we get a more concentrated density on σ^2.

12.2 A Stopping Rule for Process Optimization Based on the Conjugate Prior Regression Model

Gilmour and Mead [61] present an interesting Bayesian approach to determine if it is worth performing more experiments when the goal is to optimize a response. The answer to such question evidently provides a stopping criterion

for a sequence of designed experiments. Suppose we have k controllable factors and a single response we wish to maximize. Our goal is to find $\boldsymbol{x}_{max}$, the settings that maximize the response. Suppose $y = Y(\boldsymbol{x}) + \varepsilon$, where the ε's form a sequence of i.i.d. $N(0, \sigma^2)$ random variables and $Y(\boldsymbol{x})$ represents the expected response.

Gilmour and Mead propose to estimate the quantity

$$L(\widehat{\boldsymbol{x}_{max}}) = Y(\boldsymbol{x}_{max}) - Y(\widehat{\boldsymbol{x}_{max}}) \tag{12.5}$$

the difference between the maximum *true* expected response and the expected response at the *estimated* maximizer. If this quantity, the *expected gain*, is small, it implies our current estimate of the optimum is providing a good solution, and this implies that conducting further experiments is not be warranted. The problem, evidently, is that we do not know $Y(\boldsymbol{x}_{max})$, the true response maximal value. The proposal of Gilmour and Mead is to estimate it from the posterior distribution of the parameters as follows.

Suppose the form of the expected response is a quadratic polynomial:

$$E[Y(\boldsymbol{x})] = b_0 + \boldsymbol{b}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}. \tag{12.6}$$

Then, we know that

$$\boldsymbol{x}_{max} = -\frac{1}{2}\boldsymbol{B}^{-1}\boldsymbol{b} \quad \text{and} \quad \widehat{\boldsymbol{x}_{max}} = -\frac{1}{2}\widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{b}}. \tag{12.7}$$

Therefore, substituting $\boldsymbol{x}_{max}$ into (12.6) we get

$$\begin{aligned} Y(\boldsymbol{x}_{max}) &= b_0 - \frac{1}{2}\boldsymbol{b}'\boldsymbol{B}^{-1}b + \frac{1}{4}\boldsymbol{b}'\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{b} \\ &= b_0 - \frac{1}{4}\boldsymbol{b}'\boldsymbol{B}^{-1}\boldsymbol{b}. \end{aligned}$$

Similarly, substituting $\widehat{\boldsymbol{x}_{max}}$ in (12.6) we obtain

$$Y(\widehat{\boldsymbol{x}_{max}}) = b_0 - \frac{1}{2}\boldsymbol{b}'\widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{b}} + \frac{1}{4}\widehat{\boldsymbol{b}}'\widehat{\boldsymbol{B}}^{-1}\boldsymbol{B}\widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{b}}.$$

Therefore, the difference gives

$$L(\widehat{\boldsymbol{x}_{max}}) = -\frac{1}{4}(\boldsymbol{b}'\boldsymbol{B}^{-1}\boldsymbol{b} - 2\boldsymbol{b}'\widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{b}} + \widehat{\boldsymbol{b}}'\widehat{\boldsymbol{B}}^{-1}\boldsymbol{B}\widehat{\boldsymbol{B}}^{-1}\widehat{\boldsymbol{b}}). \tag{12.8}$$

Gilmour and Mead [61] suggest to simulate $\boldsymbol{b}$ and $\boldsymbol{B}$ from the posterior distribution of the parameters (assuming a conjugate prior[6]), $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$. The

[6]Gilmour and Mead used a conjugate model assuming the variance σ^2 is known; in the example below we use a conjugate prior on all parameters, including σ^2.

"hatted" parameters are least squares estimates. From each generated value of the parameters a value of $L(\widehat{\boldsymbol{x}_{max}})$ is computed. The set of all generated values of $L(\widehat{\boldsymbol{x}_{max}})$ approximate its posterior distribution. From this posterior distribution, quantities of interest, such as the median of the expected gain or some percentile like $\Pr(L > l_0)$ can be easily obtained. If the current data and prior indicate that the estimated optimal operating conditions are far from the optimum, this is evidence there is a potential benefit to be obtained from performing additional experiments.

Example. Gilmour and Mead [61] demonstrate their approach with the following experiment, taken from Montgomery [110]. The experimental data is shown on Table 12.2, and the analysis will be conducted assuming first that only the part of the experiment labeled "experiment 1" is run, then the runs

Table 12.2. Experimental design used to illustrate Gilmour and Mead's stopping criterion

Exp. #	x_1	x_2	x_3	Response ***(y)***
1,2,3	-1	-1	-1	66
1,2,3	-1	1	1	60
1,2,3	1	-1	1	70
1,2,3	1	1	-1	100
1,2,3	0	0	0	113
1,2,3	0	0	0	100
2,3	-1	-1	1	70
2,3	-1	1	-1	78
2,3	1	-1	-1	80
2,3	1	1	1	75
2,3	0	0	0	118
2,3	0	0	0	88
3	-1.682	0	0	100
3	1.682	0	0	80
3	0	-1.682	0	68
3	0	1.682	0	63
3	0	0	-1.682	65
3	0	0	1.682	82
3	0	0	0	100
3	0	0	0	85

labeled experiment #2 are conducted, etc. Suppose that initially, the prior on the parameter estimates has a mean of $\boldsymbol{\beta}_0' = (100, 0, 0, 0, 0, 0, 0, -6, -6, -6)$, with prior variance equal to $\sigma_0^2 = 10$ and $\boldsymbol{A}_0 = \boldsymbol{I}$. We will repeat the analysis for various values of the degrees of freedom, v_0, with smaller values of v_0 providing flatter (less informative) priors.

Drawing 10,000 simulated values of the posterior distribution of $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$, and substituting these into expression (12.8) provides and approximation to the posterior distribution of the potential gains, $L(\widehat{\boldsymbol{x}_{max}})$. From these, it is easy to compute useful quantities. For example, since the distribution of $L(\widehat{\boldsymbol{x}_{max}})$ is very skewed and non-negative, estimates of its mean will be hard to obtain. Much better estimates are obtained, for a relative low computational cost, for the median of $L(\widehat{\boldsymbol{x}_{max}})$, and some percentile, for example, $P(L(\widehat{\boldsymbol{x}_{max}}) > 5)$. Table 12.3 shows the results, utilizing different priors. The Matlab program `ComputeGainConjugate.m` (see Section 12.10) was used to carry out the computations. It is important to point out that those realizations of $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$ that did not result in a concave function $Y(\boldsymbol{x})$ were discarded, since in those cases the maximum is not defined[7].

The results in the table clearly show that we gain more from further experiments the less we know initially, i.e., the median of $L(\widehat{\boldsymbol{x}_{max}})$ and $P(L(\widehat{\boldsymbol{x}_{max}}) > 5)$ both increase as the prior distribution gets flatter (i.e., as v_0 decreases). Thus, if we are very confident in our prior, the $L(\widehat{\boldsymbol{x}_{max}})$ criterion indicates there is very little we will gain from experiments, and this is according to our intuition. The problem, as always in Bayesian analysis, is if our prior is adequate. A sensitivity analysis on the prior, similar to that shown in Table 12.3 is suggested to determine the "robustness" of our conclusion of whether or not to stop experimenting. In this example, it seems that after the first experiment it is still worth performing further experiments, but after the third experiment (which completes the central composite design) there seems to be not much more we could gain from further experiments.

Note how in the last column we show the classical non-informative (Jeffreys) case, for which the $\boldsymbol{X}'\boldsymbol{X}$ is invertible only after the 3rd experiment, when $N > p$. For that column, $v_0 = -p$, and in addition, $\boldsymbol{A}_0 = \boldsymbol{\beta}_0 = \sigma_0^2 = 0$.

[7]This follows the recommendation in Gilmour and Mead [61]. It could be argued that non-concave cases, if exist, point out to the possibility that the function is a saddle with non-negligible probability, and this should be taken into account perhaps utilizing the Ridge Analysis technique. However, we will not show the details of such approach here.

Table 12.3. Results of Gilmour and Mead's technique. Table shows median of $L(\widehat{\boldsymbol{x}_{max}})$ and $P(L(\widehat{\boldsymbol{x}_{max}}) > 5)$ obtained from 10,000 simulations of a conjugate prior Bayesian regression model, assuming σ^2 unknown, using increasingly flatter priors

Exp. #	$v_0 = 1000$	100	10	1	-10
1	0.6511, 0.0288	0.7619, 0.0436	1.4793, 0.1747	2.4411, 0.3009	N/A, N/A
2	0.3567, 0.0112	0.5379, 0.0365	1.2224, 0.1488	1.6445, 0.2214	N/A, N/A
3	0.0881, 0.0004	0.1891, 0.0046	0.5475, 0.0413	0.7295, 0.0681	1.5783, 0.1724

12.3 Bayesian Multivariate Regression and its Use in Process Optimization and Robust Parameter Design

We now extend the univariate Bayesian regression approach of the previous section to the case there are q responses. The model is:

$$\underbrace{\boldsymbol{y}_k}_{(q\times 1)} = \underbrace{\boldsymbol{B}}_{(q\times p)}\underbrace{\boldsymbol{x}_k}_{(p\times 1)} + \underbrace{\boldsymbol{u}_k}_{(q\times 1)} \quad k = 1, 2, \ldots, N$$

$$\boldsymbol{u}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \quad k = 1, 2, \ldots, N$$

Let $\boldsymbol{X}$ be the $N \times p$ design matrix formed by stacking the N $\boldsymbol{x}_k$'s, one per run. The model can then be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}' + \boldsymbol{U}$, a $N \times q$ matrix of responses made up by the N $\boldsymbol{y}_k$'s, one per row. Matrix $\boldsymbol{U}$ is also $N \times q$, and formed similarly from the $\boldsymbol{u}_k$'s.

We want to predict a new observation $\tilde{\boldsymbol{y}}$ at new settings $\boldsymbol{w}$. With the predictive density we will be able to

$$\max_{\boldsymbol{w}\in\mathcal{R}} \; P(\tilde{\boldsymbol{y}} \in A|\text{data}, \boldsymbol{w})$$

where A is some specification region for the responses [125]. The derivation of the predictive density for the multivariate case is analogous to the univariate case. A summary of the results follows.

Under the usual (i.e., Jeffreys) non-informative prior on $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, namely,

$$p(\boldsymbol{B}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{B})p(\boldsymbol{\Sigma})$$

where $p(\boldsymbol{B}) \propto$ constant, and

$$p(\boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{(q+1)/2}}$$

so that

$$p(\boldsymbol{B}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{(q+1)/2}},$$

the posterior predictive distribution of $\tilde{\boldsymbol{y}}$ is:

$$P(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) \propto \frac{1}{(1 + \frac{1}{v}(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{B}}\boldsymbol{w})'\boldsymbol{H}(\tilde{\boldsymbol{y}} - \hat{\boldsymbol{B}}\boldsymbol{w}))^{-\frac{v+q}{2}}}$$

where

$$\begin{aligned} v &= N - p - q + 1 \\ \boldsymbol{H} &= \frac{v\,\boldsymbol{S}^{-1}}{1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}} \text{ is a } q \times q \text{ matrix,} \\ \widehat{\boldsymbol{B}}' &= [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}]' \;\Leftrightarrow\; \widehat{\boldsymbol{B}} = \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ \text{and } \boldsymbol{S} &= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}) \text{ is also a } q \times q \text{ matrix.} \end{aligned}$$

In other words,

$$\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w} \sim t_v(\hat{\boldsymbol{B}}\boldsymbol{w}, \boldsymbol{H}^{-1})$$

which is a non-central multivariate student t distribution (see Appendix D) with:

$$E[\tilde{\boldsymbol{Y}}|\text{data}, \boldsymbol{w}] = \hat{\boldsymbol{B}}\boldsymbol{w}$$

and

$$\text{Var}[\tilde{\boldsymbol{Y}}|\text{data}, \boldsymbol{w}] = \frac{v}{v-2}\boldsymbol{H}^{-1} \quad (v > 2 \;\Rightarrow\; N > p + q + 1).$$

With this predictive posterior density we can compute, for any $\boldsymbol{w}$:

$$P(\tilde{\boldsymbol{y}} \in A|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) = \int_A P(\tilde{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{w}) d\tilde{\boldsymbol{y}}.$$

This integral turns out to be easy to compute, as the multivariate t distribution is easy to simulate using Monte Carlo methods:

1 simulate $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{H}^{-1})$, a $p \times 1$ vector;

2 simulate $c \sim \chi^2_v$;

3 let $\tilde{\boldsymbol{y}} = \boldsymbol{u}\sqrt{\frac{v}{c}} + \widehat{\boldsymbol{B}}\boldsymbol{w}$

4 iterate to get a sample (multivariate) histogram

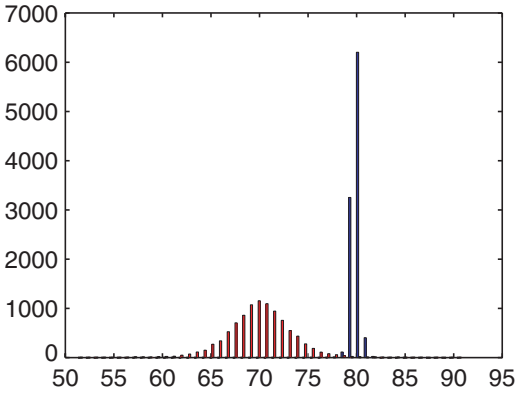


Figure 12.2. Histograms of the marginal densities of the viscosity and yield (left and right, respectively) responses, from 10,000 samples at point $x_1 = x_2 = 0$

Example. Computation of a multivariate predictive probability. Consider the chemical experiment data in the previous section. Assume we are interested this time in modelling the yield and viscosity responses simultaneously. In particular, suppose that the process engineer is interested in computing $P(\text{yield} > 80 \text{ and } 62 < \text{ viscosity } < 68)$ at the origin ($x_1 = x_2 = 0$). The Matlab program `SimulateMultiPred.m` (see Section 12.10) computes this probability for this problem, using the Monte Carlo simulation just described. The estimated predictive probability is equal to 0.1552, based on 10,000 simulation draws. Figure 12.2 shows histograms of the marginal distributions of viscosity ($\tilde{y}_2$) and Yield ($\tilde{y}_1$). As it can be seen, there is no much area in the viscosity histogram from 62 to 68. Since there are two responses only, a nice way to visualize the reason for the low joint probability is to prepare a scatter plot of the simulated y_1 and y_2 (Figure 12.3). The estimated probability is given by the draws that fell in the boxed region divided by 10,000, the total number of draws. ■

12.3.1 Advantages of the Bayesian Predictive Approach

Using the Bayesian predictive density for process optimization offers several advantages over classical approaches [125]:

1 It considers uncertainty in parameters. Classical methods for process optimization provide a single point estimate where the process should be run. This, however, neglects the variability of the parameters estimates.

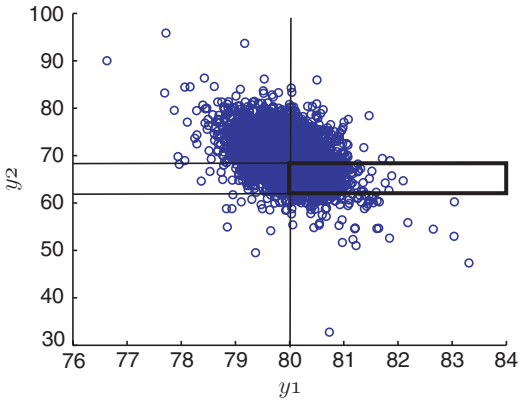


Figure 12.3. Simulated Bivariate predictive density for yield (y_1) and viscosity (y_2) obtained from 10,000 samples at point $x_1 = x_2 = 0$. Boxed region corresponds to specification region $A = \{(y_1, y_2) : y_1 > 80, 62 \leq y_2 \leq 68\}$

If a different experiment is run and a model is fitted using OLS, then a different optimum is obtained. If a confidence region on the optimal point is provided[8] (something not done much in practice), this can only be interpreted in the classical sense of being a region that would result after repeated sampling and optimization, but *not* as a region that contains the optimum with some probability;

2 It considers correlation between responses. One of the central problems in the optimization of multiple response functions based on regression models is that the responses may be correlated, and it is not clear how to account for this correlation. In the classical approach, the correlation can be accounted when fitting the models, but then the optimization step neglects the correlation. The Bayesian predictive approach considers the whole multivariate distribution of the responses, hence it takes into account explicitly the correlation between the responses *at the optimization step*;

3 It is "objectively" Bayesian. The predictive densities presented here were based on non-informative priors. This means that no strong priors are used. If necessary, however, prior elicitation methods for regression models (see [73]) can be used (Kadane et al.'s elicitation is based on asking questions

[8]See [126].

about the responses, not about the parameters, making their procedure much more practical);

4 It can tell how "sweet" a "sweet spot", obtained through controur plots, is. It is a widely accepted practice to draw contours plots of OLS-fitted responses and use the plots to locate a "sweet spot" where analysts think the process can be run. However, the *probability* of observing future responses in that "spot" is not known. The predictive approach can be used to compute such probability. Not surprisingly, for multiple responses these probabilities can be very low, given that the contour plots are based on models for the mean of each response;

5 It can be used for more general optimization problems, such as:

$$\max_{\boldsymbol{w} \in \mathcal{R}} P(g(\tilde{\boldsymbol{Y}})|\text{data}, \boldsymbol{w})$$

for any scalar function $g(\bullet)$. For example: $P(D(\tilde{\boldsymbol{Y}}) > d|\text{data}, \boldsymbol{w})$ gives the probability that the overall desirability function exceeds some value d (recall that the desirability method, widely used in RSM, does not provide a measure of how likely it is to have a "desirable" response at the optimum operating conditions). Another example is $P(Q(\tilde{\boldsymbol{Y}}) < r|\text{data}, \boldsymbol{w})$, the probability that a quadratic cost function does not exceed some value r.

6 It allows to do a "pre-posterior" analysis. If the maximum posterior probability is low for a given process, it could be due to a) very stringent constraints on the responses, and/or due to b) poor model fit. A "preposterior analysis" (see Peterson [125]) allows us discern between these two situations, determining if further experiments would improve $P(g(\tilde{\boldsymbol{Y}})|\text{data}, \boldsymbol{w})$ through better models.

7 It is easy to add noise variables, providing a Bayesian approach to multivariate Robust Parameter Design.

Example. Evaluation of the probability of conformance in the desirability method. Consider Derringer and Suich's [45] classic tire tread compound problem (Table 4.7). In this example, there are $q = 4$ responses and $p = 3$ controllable factors (no noise factors). The DOE used was a CCD, with $N = 20$ runs. Full quadratic polynomials were fitted to all four responses using OLS.

The solution using the authors' desirability method, based on

$$\max_{\boldsymbol{w}} D(\boldsymbol{w}) = \left(\prod_{j=1}^{q} d_j(Y_j(\boldsymbol{w})) \right)^{1/q}$$

gives the point $\boldsymbol{w}^* = (-0.050, 0.145, -0.868)'$. Figure 12.4 shows the desirability function contour plot.

Using the predictive density approach with region:

$$\begin{aligned} A &= \{120 < y_1 < \infty, 1000 < y_2 < \infty, 400 < y_3 < 600, \\ &\quad 60 < y_4 < 75\} \end{aligned}$$

(same as specified in Derringer and Suich) we get that $P(\tilde{\boldsymbol{Y}} \in A | \text{data}, \boldsymbol{w}^*) = 0.47$. This means that with current DOE data and regression models, chances

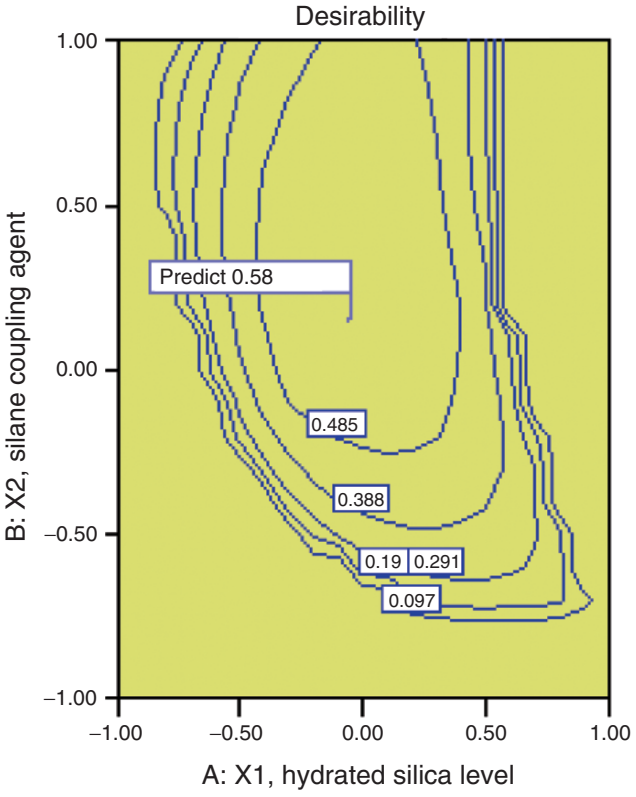


Figure 12.4. Desirability function contour plot, Derringer and Suich tire tread compound example

Table 12.4. Marginal probabilities of satisfying the specifications, Derringer and Suich example

	One rep.	**2 reps.**	**3 reps.**
$P(120 < \tilde{y_1} < \infty \vert \text{data}, \boldsymbol{x}^*)$	0.68	0.83	0.90
$P(1000 < \tilde{y_2} < \infty \vert \text{data}, \boldsymbol{x}^*)$	0.67	0.82	0.89
$P(400 < \tilde{y_3} < 600 \vert \text{data}, \boldsymbol{x}^*)$	0.99	1.0	1.0
$P(60 < \tilde{y_4} < 75 \vert \text{data}, \boldsymbol{x}^*)$	0.99	1.0	1.0

are *very low* (lower than flipping a coin) that the "optimum" $\boldsymbol{w}^*$ will satisfy the design constraints. A preposterior analysis can be conducted by pretending we have more data than we actually have. For example, we could repeat the $\boldsymbol{X}$ matrix and change the degrees of freedom accordingly, to simulate the effect of more data. This assumes we have *repeatability* of the data in the sense that the new data being implicitly simulated in this way would result in exactly the same $\widehat{\boldsymbol{\beta}}$ and s^2 least square estimates as the old data[9]. However, if for this very optimistic scenario the probability of meeting the constraints is still too low, we know that the constraints are too demanding, and no further experiments are justified unless the "design" of the product or process under study changes. For the tire tread compound example, we have that with two replicates we get $P(\tilde{\boldsymbol{y}} \in A \vert \text{data}, \boldsymbol{w}^*) = 0.68$ and we three replicates $P(\tilde{\boldsymbol{y}} \in A \vert \text{data}, \boldsymbol{w}^*) = 0.80$. Therefore, it seems that models can benefit from running further experiments. The marginal probabilities of conformance of each response are shown in Table 12.4. It is clear that the problem is with the fit of the first two responses. Note how the point estimate at the optimum conditions obtained with the desirability method give response values well inside the specification region A: $\hat{y_1} = 129.5, \hat{y_2} = 1300, \hat{y_3} = 465$, and $\hat{y_4} = 68$. ■

12.3.2 Inability of the Frequentist Approach to Provide a Probability of Conformance

In classical (frequentist) regression, a *prediction interval* is constructed from the pivot

$$\frac{e}{\sqrt{\widehat{\text{Var}}(e)}} = \frac{\tilde{y}|\boldsymbol{w} - \widehat{y}|\boldsymbol{w}}{s\sqrt{1 + \boldsymbol{w}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{w}}} \sim t_{N-p}.$$

[9]Recall that the data enters the posterior predictive density only through the $\widehat{\boldsymbol{\beta}}$ and s^2 estimates.

where the point estimate for the next observation $\tilde{y}$ at point $\boldsymbol{w}$ is $\widehat{y}|\boldsymbol{w} = \boldsymbol{w}'\widehat{\boldsymbol{\beta}} = \widehat{E}[\tilde{y}|\boldsymbol{w}]$. That is, we take the prediction error, $\tilde{y}|\boldsymbol{w} - \widehat{y}|\boldsymbol{w}$, divided by its estimated standard error, which leads to a Student t distribution. This is used to setup a valid prediction interval on $\tilde{y}|\boldsymbol{w}$, but notice that we *still* have that the distribution of the future observation is simply $\tilde{y}|\boldsymbol{w} \sim N(\boldsymbol{w}'\boldsymbol{\beta}, \sigma^2)$. The t distribution so obtained is for the prediction error, where $\widehat{y}|\boldsymbol{w}$ is random. The distribution of the pivot coincides with the predictive density of $\tilde{y}$ under a non-informative prior. However, the *fundamental* difference is that in the Bayesian approach the probability measure is associated to $\tilde{y}|\boldsymbol{w}$, not to $\widehat{y}|\boldsymbol{w} = \boldsymbol{w}'\widehat{\boldsymbol{\beta}}$, which in the Bayesian setting is simply a constant. But a probability measure on $\tilde{y}$ that relates to our model is what we need to compute "reliabilities" (i.e., probabilities of conformance) of the form $P(\tilde{y}|\text{data}, \boldsymbol{w} \in A)$, where A is a region defined by product specifications, as in the example. Therefore, there is *no* way to compute a probability of conformance for $\tilde{y}$ from the prediction intervals (or regions, in case $\tilde{y}$ is multivariate), in the classical-frequentist approach. In the frequentist approach, the best we can do is compute the probability of conformance of $\widehat{y}$, the *predicted* response (see Section 9.5 for a specific proposal on how to compute the mean and variance of $\widehat{y}$ in RPD taking into account the estimation variability). From a frequentist perspective, $\tilde{y}$, the *next* or future response value is simply a normal random variable defined by the assumed regression model, and has the same distribution as any other observation y_i has before experimental run i is conducted.

12.4 A Bayesian Approach to Robust Parameter Design

In this section we return to the Robust Parameter Design problem, discussed in Chapter 9 from a variety of frequentist approaches. Here we look at some of the same problems as in Chapter 9 but from a Bayesian perspective. Consider the model proposed in [27, 116], equation (9.1). Recall we classify factors as controllable ($\boldsymbol{x}$) and uncontrollable (noise) $\boldsymbol{z}$. It is assumed that during design and experimentation all factors are controllable. For optimization, we let the noise factors vary randomly according to some distribution. The Box-Jones-Myers dual response model is

$$y = \beta_o + \boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} + \boldsymbol{z}'\boldsymbol{\gamma} + \boldsymbol{x}'\boldsymbol{\Delta}\boldsymbol{z} + \varepsilon \tag{12.9}$$

where $\boldsymbol{x}$ is a $k \times 1$ vector of controllable factors, $\boldsymbol{z}$ is a $r \times 1$ vector of noise variables, and $\varepsilon \sim (0, \sigma_\epsilon^2)$.

A Bayesian approach for this problem uses the predictive densities derived earlier in this chapter with $\boldsymbol{w}' = (\boldsymbol{x}, \boldsymbol{z})$. The noise factors are integrated with respect to their assumed density, while we optimize with respect to $\boldsymbol{x}$ only:

$$p(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \mathcal{R}} \int p(\tilde{y} \in A | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{z})) p(\boldsymbol{z}) d\boldsymbol{z}.$$

If $p(\boldsymbol{x})$ is large, we have obtained a robust solution. Otherwise, we can analyze the reason for a low probability using the preposterior approach sketched before. The following example illustrates this procedure.

Example. Addition of noise factors: Bayesian Robust Parameter Design, single response. Myers and Montgomery [117] give a Semiconductor manufacturing process in which there is a single response, 2 controllable factors, and 2 noise factors. A CCD was run (no axial runs were used in the noise factors). The model (12.9) was fitted to the data giving:

$$\begin{aligned} \widehat{Y}(\boldsymbol{x}, \boldsymbol{z}) \quad = \quad & 30.37 - 2.9x_1 - 4.13x_2 + 2.6x_1^2 + 2.18x_2^2 + 2.87x_1x_2 \\ & + 2.73z_1 - 2.33z_2 + 2.33z_3 + 0.27x_1z_1 + 0.89x_1z_2 \\ & + 2.58x_1z_3 + 2.01x_2z_1 - 1.43x_2z_2 + 1.56x_2z_3 \end{aligned}$$

Assuming that $E[\boldsymbol{z}] = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_z = \boldsymbol{I}_3$, then:

$$E_z[\widehat{Y}(\boldsymbol{x}, \boldsymbol{z})] = 30.37 - 2.92x_1 - 4.13x_2 + 2.60x_1^2 + 2.18x_2^2 + 2.87x_1x_2$$

and[10]

$$\text{Var}_z[\widehat{Y}(\boldsymbol{x}, \boldsymbol{z})] = 19.26 + 3.20x_1 + 12.45x_2 + 7.52x_1^2 + 8.52x_2^2 + 2.21x_1x_2.$$

We are told that the manufacturer wants $E_z[\widehat{Y}(\boldsymbol{x}, \boldsymbol{z})] < 30$ and $\sqrt{\text{Var}_z[\widehat{Y}(\boldsymbol{x}, \boldsymbol{z})]} < 5.5$. The classical approach finds a feasible region by superimposing the contour plots of the responses. Figure 12.5 shows the "sweet spot" where it seems the constraints will be met.

Let us apply the Bayesian predictive approach to a point $\boldsymbol{x}$ within the "sweet spot", say $\boldsymbol{x}' = (0, 0)$. At this point, $p(0 < \tilde{y} < 30 | \text{data}, \boldsymbol{x}) \simeq 0.48$. If we instead maximize with respect to the controllable factors, we obtain $\max_{\boldsymbol{x}} p(0 < \tilde{y} < 30 | \text{data}, \boldsymbol{x}) = 0.61$ at $\boldsymbol{x}' = (0.1, 0.8)$. A histogram of the predictive density of $\tilde{y}$ at this point is shown in Figure 12.6, together with a contour plot of the predictive density for all $\boldsymbol{x}$.

[10]Readers will notice that this variance is the *biased* estimator of the variance that was discussed in Chapter 9, $\text{Var}_{\boldsymbol{z}}^{(b)}[\widehat{y}(\boldsymbol{x}, \boldsymbol{z})]$. We use this estimator to illustrate how the "sweet spot" approach, with the usual variance estimator, performs when evaluated using a Bayesian formulation.

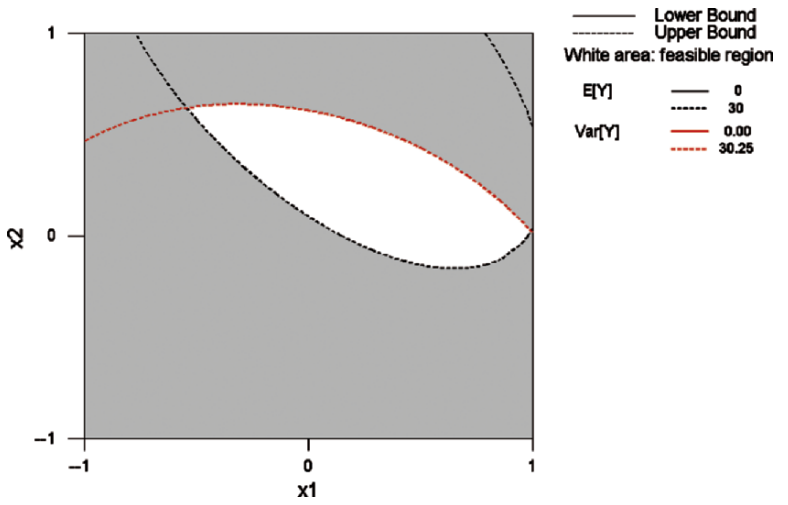


Figure 12.5. Contour plots of mean and variance responses, semiconductor manufacturing example

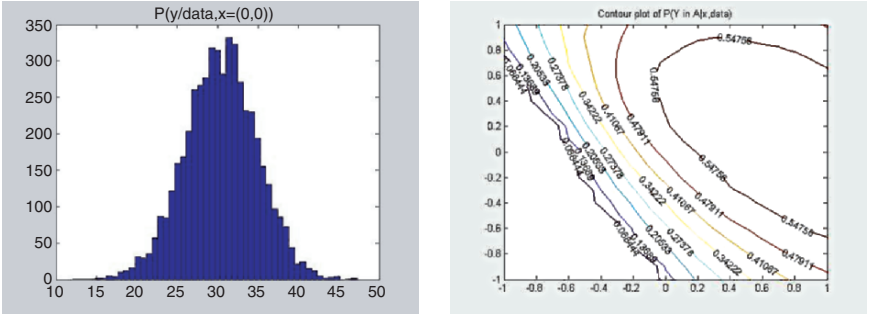


Figure 12.6. Left: Histogram of the predictive distribution at $(0.1, 0.8)$; right: contour plot of $p(\tilde{y}|\text{data}, \boldsymbol{x})$, semiconductor manufacturing example

In this case, a preposterior analysis [125] does not show a significant increase in $p(\boldsymbol{x})$ as the number of replications increases, indicating the problem is that the constraints are unrealistic. This is evident from looking at the histogram in Figure 12.6, where it can be seen that $p(0 < \tilde{y} < 30|\text{data}, \boldsymbol{x} = (0.1, 0.8)) \simeq 0.6$. Thus, under the current process conditions and specifications, no robust solution is possible. For example, if we change the specifications to $A =$

$\{0 < \tilde{y} < 35\}$ instead, $\max_{\boldsymbol{x}} P(0 < \tilde{y} < 35|\text{data}, \boldsymbol{x}) = 0.867$ is obtained at (0.20, 0.0). We note from the left figure in Figure 12.6, that for the current specification $A = \{0 < \tilde{y} < 30\}$, no benefits will result from reducing $\boldsymbol{\Sigma}_z$ either, since process centering is more important. ■

Example. Bayesian Robust Parameter Design, multiple responses. Miro et al. [106] consider the Bayesian optimization of a High Pressure Liquid Chromatography (HPLC) process, critical in the pharmaceutical industry. It is of interest to study the potential effect that two controllable factors, temperature and pH, and one noise factor, the % IPA, have on four responses: Rs, RunTime, S/N Ratio, and Tailing. A Box-Behnken design on the three factors was run with three center points. Second order polynomials were fit using OLS with very good results ($R^2 > 0.98$). Assume the % IPA varies as a $N(0, 0.01)$ random variable during manufacturing, and let the specification region of the four responses be

$$A = \left\{ y_{Rs} \geq 1.8,\ y_{Time} \leq 15,\ y_{S/N} \geq 300,\ 0.75 \leq y_{Tail} \leq 0.85 \right\}.$$

The Bayesian optimization results in $w^* = [Temp^*, pH^*]' = [0.4822, 1.0]'$ for $\max P(\tilde{\boldsymbol{Y}} \in A|\text{data}, \boldsymbol{w}^*) = 0.9622$. Figure 12.7 shows a contour and 3D plots of the posterior probability of conformance. As it can be seen, there is a "ridge" on the function along the x_2 (pH) controllable factor, which indicates that only the temperature has a significant effect on the performance of the process. ■

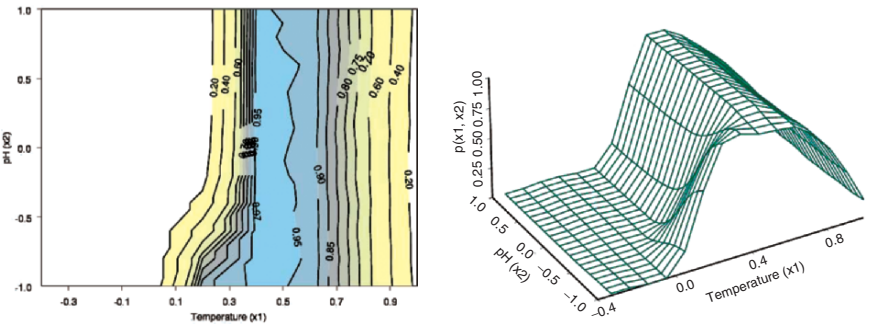


Figure 12.7. Left: Contour plot of the posterior probability of conformance to specifications, HPLC example. Right: 3D plot. Source: [106]

12.5 A Connection with Process Control**

There are interesting connections[11] between the Bayesian regression model presented here and process control. Among others, Press [129] and Zellner [162] treat this matter in detail.

Consider a multivariate "target is best case" with $\boldsymbol{T}$ a vector of targets for the q responses. Suppose there is a quadratic off-target cost and a quadratic cost of adjusting the process. The assumed predicted loss is:

$$L(\boldsymbol{y}_{n+1}) = (\boldsymbol{y}_{n+1} - \boldsymbol{T})'\boldsymbol{G}(\boldsymbol{y}_{n+1} - \boldsymbol{T}) + (\boldsymbol{x}_{n+1} - \boldsymbol{x}_n)'\boldsymbol{J}(\boldsymbol{x}_{n+1} - \boldsymbol{x}_n)$$

where, following common practice in control applications, we write $\boldsymbol{y}_{n+1}$ for the next observation of $\boldsymbol{y}$, and we use $\boldsymbol{x}_{n+1}$ to denote the next settings for the controllable factors. $\boldsymbol{G}$ and $\boldsymbol{J}$ are user-defined square matrices that represent the off-target and adjustment costs. Then the expected predicted loss is:

$$E[L(\boldsymbol{y}_{n+1})] = \int L(\boldsymbol{y}_{n+1})p(\boldsymbol{y}_{n+1}|data, \boldsymbol{w}_{n+1})d\boldsymbol{y}_{n+1}.$$

Suppose we can partition the vector of factors as $\boldsymbol{x}' = [\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(2)}]$, with q_1 controllable factors and q_2 uncontrollable factors. Likewise, partition $\widehat{\boldsymbol{B}}$, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $\boldsymbol{J}$ accordingly.

Our goal is to

$$\min_{\boldsymbol{x}^{(1)}} \; E[L(\boldsymbol{y}_{n+1})]$$

i.e., minimize the expected cost with respect to the controllable factors. Under a noninformative prior on the parameters, this problem turns out to have a closed-form solution given by ([129, 162]):

$$\begin{aligned}
\boldsymbol{x}_{n+1}^{(1)^*} \;=\; & \left[\widehat{\boldsymbol{B}^{(1)}}'\boldsymbol{G}\widehat{\boldsymbol{B}^{(1)}} + \frac{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{G})}{(v-2)(N-p)}\boldsymbol{D}^{(11)} + \boldsymbol{J}^{(11)}\right]^{-1} \\
& \times[\widehat{\boldsymbol{B}^{(1)}}'\boldsymbol{G}(\boldsymbol{T} - \widehat{\boldsymbol{B}^{(2)}}\boldsymbol{x}_{n+1}^{(2)}) - \frac{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{G})}{(v-2)(N-p)}\boldsymbol{D}^{(12)}(\boldsymbol{x}_{n+1}^{(2)} \\
& + \boldsymbol{J}^{(11)}\boldsymbol{x}_n^{(1)}) + \boldsymbol{J}^{(12)}(\boldsymbol{x}_n^{(2)} - \boldsymbol{x}_{n+1}^{(2)})]
\end{aligned}$$

The value of this formulation is that we can think of the uncontrollable factors as "noise factors". Therefore, we can measure, forecast or simulate the future values of the uncontrollable factors, $\boldsymbol{x}_{n+1}^{(2)}$ and use the expression above

[11] This section can be skipped on a first reading without loss of continuity.

to determine the optimal setting for the controllable factors, in a manner similar to Robust Parameter Design.

12.6 Model-Robust Process Optimization**

Suppose[12] $M_1, M_2, \ldots$ are several models of the form:

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

that fit the data "well". We wish to

$$\max_{\boldsymbol{x}^* \in \mathcal{R}} p(L < \tilde{y} < U \mid \text{data}, \boldsymbol{x}^*)$$

but we wish to account for *all* models. Optimizing each model separately may result in different, conflicting optimums. Bayesian model averaging can be used to account for all models. In this approach, posterior probability densities of the response are conditioned on a particular assumed model M_i. These probabilities can then be combined, or averaged, via the posterior probabilities of each potential model being true. In order to do this, define the model-average (posterior) predictive density as

$$\text{MAP } = p(\tilde{y}|\boldsymbol{x}, \boldsymbol{y}) = \sum_{\text{all } i} p(\tilde{y}|\boldsymbol{x}, \boldsymbol{y}, M_i)\ p(M_i|\ \boldsymbol{y}).$$

where $p(M_i|\boldsymbol{y})$ is the posterior probability of each model, which are explained below. If we have the means of computing the MAP, then we can maximize its integral:

$$\begin{aligned} &\max_{\boldsymbol{x} \in \mathcal{R}} \int_L^U p(\tilde{y}|\boldsymbol{x}, \boldsymbol{y})\ d\tilde{y} \\ &= \max_{\boldsymbol{x} \in \mathcal{R}} \sum_{\text{all } i} \int_L^U p(\tilde{y}|\ \boldsymbol{x}, \boldsymbol{y}, M_i)\ d\tilde{y}\ \ p(M_i|\ \boldsymbol{y}) \end{aligned}$$

Rajagopal and del Castillo [131] adopted the model priors $\pi_j, \quad j = 1, 2, \ldots, k$ described by Meyer and Box [99]. These are based on the prior probability of each factor i being active. Let f_i denote be the number of active factors in model i.

Then, assume that *a priori*,

$$p(M_i) = \prod_{j \in M_i} (\pi_j) \prod_{j' \notin M_i} (1 - \pi_{j'})$$

[12]This section is based on Rajagopal and del Castillo [131] and presents material at a relatively more advanced level.

where "$j \in M_i$" means factor j is active in model i, etc. Note that if $\pi_j = \pi$ for all $j = 1, 2, \ldots, k$, then $p(M_i) = \pi^{f_i}(1-\pi)^{k-f_i}$.

With this prior, the *model posterior* is obtained from Bayes' theorem:

$$p(M_i|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|M_i)\ p(M_i)}{\sum_{\text{all } i} p(\boldsymbol{y}|M_i)\ p(M_i)}$$

where the "marginal likelihood" is

$$p(\boldsymbol{y}|M_i) = \int\int p(\boldsymbol{y}|M_i, \sigma^2, \boldsymbol{\beta}_i)\ p(\sigma^2, \boldsymbol{\beta}_i|M_i) d\boldsymbol{\beta}_i d\sigma^2$$

and where the likelihood is

$$p(\boldsymbol{y}|M_i, \sigma^2, \boldsymbol{\beta}_i) \propto \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta}_i)'(\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\beta}_i)\}.$$

Let r_i be the number of *terms* in model M_i and let t_i be the number of terms excluding the intercept. Here, $\boldsymbol{X}_i$ is a $n \times r_i$ design matrix.

To setup a prior distribution on the parameters, recall that when using OLS, $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, so it seems it makes sense to let the prior variance of each $\boldsymbol{\beta}_i$ depend on $\boldsymbol{X}_i'\boldsymbol{X}_i$. At the same time, it makes sense to include an additional parameter to allow us to "tune" how flat we want this prior. This is called a "g-prior", common in regression analysis[13]. Thus, we assume that *a priori*,

$$\boldsymbol{\beta}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_i\ \sigma^2), \quad p(\sigma^2) \propto \frac{1}{\sigma^2} \quad \text{and} \quad p(\boldsymbol{\beta}_i, \sigma^2) = p(\sigma^2)\ p(\boldsymbol{\beta}_i)$$

where

$$\Sigma_i^{-1} = (\boldsymbol{X}_i'\boldsymbol{X}_i)\boldsymbol{V}_i, \quad \text{where} \quad \boldsymbol{V}_i = \frac{1}{g}\begin{pmatrix} 0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{t_i} \end{pmatrix}$$

Note that as $g \to 0$ we get higher precision; as $g \to \infty$ we get a flat prior. Also note that for a null model (i.e., a model with only the intercept), $r_i = 1, t_i = r - 1 = 0$, so $\boldsymbol{V}_i = 0$. This implies that $\boldsymbol{\Sigma}_i^{-1} = 0$, thus, for the intercept

$$p(\beta_0) \ \propto \ \text{constant}.$$

In summary, we use a non-informative prior on the intercept and on σ^2, and a g-prior on all other β's.

The reason we use a non-informative prior only on the intercept and not on all the parameters (as in previous sections) is that since we are doing model

[13]G-priors were proposed by Zellner [161].

comparisons, a non-informative prior on all parameters will result in the null model being the most probable a posteriori. This is clearly not useful since with such a model there is no possibility of any optimization.

For the assumed model and priors, Meyer and Box [99] showed that the marginal likelihood is:

$$p(\boldsymbol{y}|M_i) \quad \propto \quad \gamma^{t_i}|\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{X}_i'\boldsymbol{X}_i|^{-1/2} S_i^{-(n-1)/2}$$

where they used the slightly different parametrization

$$\frac{g}{\gamma}\boldsymbol{V}_i = \boldsymbol{\Sigma}_i^{-1} \quad \text{or} \quad \frac{1}{\gamma}\begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{t_i} \end{pmatrix} = \boldsymbol{\Sigma}_i^{-1}$$

thus γ^2 has the same effect as g. With this, omitting the denominator in Bayes' formula, we get

$$p(M_i|\boldsymbol{y}) \;\propto\; \pi^{f_i}(1-\pi)^{k-f_i}\gamma^{-t_i}|\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{X}_i'\boldsymbol{X}_i|^{-1/2} S_i^{-(n-1)/2} \quad \text{where}$$

$$S_i = (\boldsymbol{y} - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}_i})'(\boldsymbol{y} - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}_i}) + \widehat{\boldsymbol{\beta}_i}'\boldsymbol{\Sigma}_i^{-1}\widehat{\boldsymbol{\beta}_i}$$

is the sum of squares and where the posterior point estimate of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}_i} = (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{y}.$$

Note how if $\gamma^2 \to \infty$ we get the OLS estimator.

Under the stated priors and model assumptions, Rajagopal and del Castillo [131] show that

$$p\left(\frac{\tilde{y} - \boldsymbol{w}'\widehat{\boldsymbol{\beta}_i}}{\widehat{\sigma}_i\sqrt{1 + \boldsymbol{w}'(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{X}_i\boldsymbol{X}_i)^{-1}\boldsymbol{w}}} < t \mid M_i, \boldsymbol{w}, \boldsymbol{y}\right) = \frac{1}{2}\left[1 + I_{t^2/(v+t^2)}\left(\frac{1}{2}, \frac{v}{2}\right)\right] \tag{12.10}$$

where $I_z(a,b)$ is the incomplete beta function, $v = n-1$, and $\widehat{\sigma}^2 = S_i/(n-1)$. With this and $p(M_i|\boldsymbol{y})$ we can easily compute the cumulative MAP and get an optimal, model-robust solution.

12.6.1 Selection of the Hyperparameters and Family of Models

In the previous formulation, there are two hyperparameters that need to be defined by the user: π and γ. Rajagopal et al. show how for a given choice of γ, the influence of π on the solutions is not significant. They used $\pi = 0.5$,

which is informative in that it implies that half of the factors are expected to be included *a priori* in each model.

To select γ, the most critical hyperparameter, these authors followed an approach proposed by Meyer and Box [99] in which γ is selected such that it minimizes $p(M_0|\boldsymbol{y})$, the posterior probability of the null model. This is, in fact, not a purist Bayesian approach but an "Empirical Bayes" approach, as the prior is data-based.

Finally, a question concerns what do we mean about "all models". A practical answer is that by all models we mean models within a family or families of models we choose from which we compute $p(M_i|\text{data})$ for all members of these families. The probabilities need to be normalized by making them add up to one.

An example illustrates this procedure. Another example (on a mixture experiment) is presented in [131].

Example. Model Robust Optimization of a semiconductor process. Consider an experiment on a chemical vapor deposition (CVD) process used in semiconductor manufacturing [36]. The goal of the experiment was to investigate the Uniformity and Stress responses. This examples illustrates the model-averaging approach on the first response. The central composite inscribed (CCI) design that was used and the experimental data are shown in Table 12.5. There are two controllable factors: Pressure and ratio of the gaseous reactants H_2 and WF_6 (denoted by H_2/WF_6). The goal was to minimize the response, as a smaller value of "Uniformity" indicates a more uniform layer being deposited on a wafer. The models considered included combinations of main effects, two-way interactions and quadratic effects. In all the models higher order effects were included only if the corresponding main effect(s) is(are) present in the model. Table 12.6 lists these models along with their least square regression statistics and posterior probabilities. The prior on the factors, π, was set at 0.5 and a value of $\gamma = 2$ minimized the posterior probability of the null model.

Models with $P(M_i|data) > 0.0254$ were used for model averaging as they accounted for 95% of the probability. Based on these models and within the region $\{-1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1\}$, the MAP was maximized for $\tilde{y} \in (-\infty, 5)$ at the point $(1.0000, -0.9198)$ yielding a maximum probability of conformance of 0.8851. The optimum values of the controllable factors

Table 12.5. Design and experimental data for the CVD process (from [36])

Coded Pressure	**Coded H_2/WF_6**	**Uniformity**
1	0	4.6
0	0	6.2
0.71	−0.71	3.4
−0.71	0.71	6.9
−1	0	7.3
0	0	6.4
−0.71	−0.71	8.6
0	−1	6.3
0.71	0.71	5.1
0	1	5.4
0	0	5

Table 12.6. Least square regression statistics and posterior probabilities for competing models for the CVD example. From [131]

$Model$	$constant$	A	B	AB	A^2	B^2	R^2	R^2_{Adj}	$S.E.$	$P(M_i\|data)$
1	1	1	1	1	0	0	0.8703	0.8148	0.6145	0.2827
2	1	1	0	0	0	0	0.7186	0.6874	0.7982	0.2396
3	1	1	1	1	1	0	0.8715	0.7858	0.6607	0.1080
4	1	1	1	1	0	1	0.8703	0.7839	0.6637	0.1053
5	1	1	0	0	1	0	0.7198	0.6498	0.8449	0.0907
6	1	1	1	0	0	0	0.7285	0.6607	0.8316	0.0671
7	1	1	1	1	1	1	0.8716	0.7431	0.7235	0.0416
8	1	1	1	0	1	0	0.7297	0.6139	0.8871	0.0254
9	1	1	1	0	0	1	0.7285	0.6122	0.8891	0.0250
10	1	1	1	0	1	1	0.7298	0.5496	0.9581	0.0098
11	1	0	0	0	0	0	0.0000	0.0000	1.4276	0.0035
12	1	0	1	0	0	0	0.0099	−0.1001	1.4974	0.0009
13	1	0	1	0	0	1	0.0099	−0.2376	1.5882	0.0003

obtained by maximizing the individual predictive densities, and the maximum value of the predictive density for the individual models for $\tilde{y} \in (-\infty, 5)$ are given in Table 12.7. It can be seen that for all the models the optimum value of x_1 is 1, but the optimum setting for x_2 can vary anywhere from -1 to 1. Figure 12.8 shows the surface plot of the cumulative posterior predictive density of the response in the region $(-\infty, 5)$ for different possible values of the control factors. In order to better understand the importance of maximizing the MAP, Table 12.8 shows the probabilities of conformance, $P(-\infty < \tilde{y} < 5)$ for various cases of the true model and the assumed model. The table shows

Table 12.7. Optimum for individual models for the CVD example. From [131]

$Model\ no.$	$P(M_i \mid data)$	x_1^*	x_2^*	z^*
1	0.2827	1	-1	0.9665
2	0.2396	1	N/A	0.8132
3	0.1080	1	-1	0.9569
4	0.1053	1	-0.9017	0.9618
5	0.0907	1	N/A	0.7776
6	0.0671	1	1	0.8477
7	0.0416	1	-0.9018	0.9464
8	0.0254	1	1	0.8178

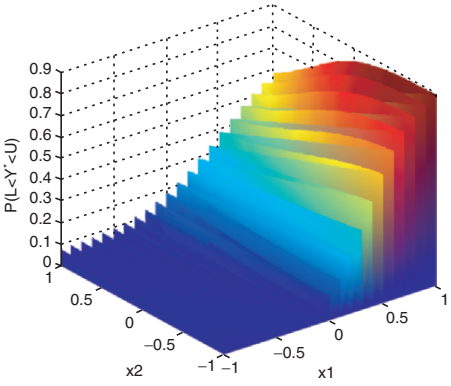


Figure 12.8. $P(\tilde{y} \in (-\infty, 5))$ as a function of x_1 and x_2 for the CVD example. From [131]

Table 12.8. Model-robustness analysis for the CVD example. Table gives $P(L \leq \tilde{y} \leq U|M_i, \mathbf{y}, x_1, \ldots, x_k)$ where M_i is the true model, evaluated at the settings $x_1, \ldots, x_k$ obtained from maximizing the probability of conformance using the assumed model. From: [131]

$Assumed\ Model \rightarrow$	1	2	3	4	5	6	7	8	MAP
$True\ Model \downarrow$									
1.00	0.9665	0.8680	0.9665	0.9648	0.8680	0.3935	0.9648	0.3935	0.9651
2.00	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132
3.00	0.9569	0.8323	0.9569	0.9540	0.8323	0.3733	0.9540	0.3733	0.9546
4.00	0.9612	0.8614	0.9612	0.9618	0.8614	0.3991	0.9618	0.3991	0.9618
5.00	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776
6.00	0.7379	0.8165	0.7379	0.7472	0.8165	0.8477	0.7472	0.8477	0.7455
7.00	0.9456	0.8318	0.9456	0.9464	0.8318	0.3749	0.9464	0.3749	0.9463
8.00	0.7043	0.7809	0.7043	0.7130	0.7809	0.8178	0.7130	0.8178	0.7114
Min	0.7043	0.7776	0.7043	0.7130	0.7776	0.3733	0.7130	0.3733	0.7114
Max	0.9665	0.8680	0.9665	0.9648	0.8680	0.8477	0.9648	0.8477	0.9651
Mean	0.8579	0.8227	0.8579	0.8598	0.8227	0.5996	0.8598	0.5996	0.8594
Std. Dev.	0.1111	0.0330	0.1111	0.1075	0.0330	0.2302	0.1075	0.2302	0.1082

the value of $P(L \leq \tilde{y} \leq U|M_i, \mathbf{y}, x_1, \ldots, x_k)$ where M_i is the true model and control factors $x_1, \ldots, x_k$ are set at their optimal values obtained from maximizing this probability using the *assumed* model. Thus, for example, if the assumed model is model 1, then the probability of conformance is maximized at the point $(1, -1)$, as shown in Table 12.7, yielding a probability of 0.9665. However, this is actually the probability of conformance only if the true model is also model 1. If, for example, it so happens that the true model is model 8, then the probability of having $\tilde{y} \in (-\infty, 5)$ is actually 0.7043 when using the solution point $(1, -1)$, obtained with the wrong model. Similarly, the last column on the table shows $P(L \leq \tilde{y} \leq U|M_i, \mathbf{y}, x_1, \ldots, x_k)$ for the true model, evaluated at the solution $x_1, \ldots, x_k$ obtained from maximizing the MAP. Based on the column statistics, it can be seen that operating at the point which maximizes the MAP has highest average probability of conformance (and among lowest std. deviation of this probabilities) compared to probabilities provided by solutions obtained by assuming single one of the competing models. The MAP also has higher minimum probability of conformance, thus it improves the worst-case scenario (worst true model). Therefore, it is seen that regardless the true process model (within the assumed family of models), the solution obtained using the model-average approach provides an operating point that gives relative high probabilities of conformance. It is in this sense that we can say the solutions obtained are *robust* to the uncertainty in the form of the true model[14].

Table 12.9 shows the sensitivity of the solution with respect to the chosen parameters γ and π. It is seen that the sensitivity of the solution to π is dependent on γ. At the value of γ chosen, the optimal controllable variables as well as the optimal predictive density are insensitive to the choice of π. ■

12.7 Model-Robust Optimization with Noise Factors

It is possible to perform model-robust optimization based on the MAP approach while at the same time protecting against variations in the noise factors. The solutions thus obtained are robust with respect to variations in the model form, uncertainty in the parameters in each model, and variation in the noise factors themselves. An additional numerical integration on the MAP is

[14]Note that models 2 and 5 are independent of the second factor, x_2 (H_2/WF_6). In the table, for the columns associated with these two models, the probabilities of conformance were evaluated at the point $(1, 0)$.

Table 12.9. Sensitivity of solution with respect to parameters π and γ for the CVD example. Source: [131]

γ	π	x_1^*	x_2^*	z^*
0.5	0.25	1	-1	0.4878
0.5	0.50	1	-1	0.5191
0.5	0.75	1	-1	0.5218
1	0.25	1	-0.9151	0.7149
1	0.50	1	-0.9154	0.7487
1	0.75	1	-0.9154	0.7621
2	0.25	1	-0.9684	0.8428
2	0.50	1	-0.9198	0.8851
2	0.75	1	-0.9198	0.9094
5	0.25	1	-0.7760	0.8438
5	0.50	1	-0.7928	0.8992
5	0.75	1	-0.7930	0.9305
10	0.25	1	-1	0.8090
10	0.50	1	-0.5696	0.8557
10	0.75	1	-0.5995	0.8967
100	0.25	1	-0.9072	0.5563
100	0.50	1	0.5146	0.7038
100	0.75	1	0.7584	0.7944

performed over the distribution of the noise factors. See Rajagopal, del Castillo and Peterson [133].

12.8 Bayesian Optimization with Mixture Models

We now consider Bayesian optimization of processes that can be described as *mixtures* of components (see Section 5.8). Berliner [8] studied a specific Bayesian optimization problem where a product is formulated as a mixture that results in an easy to solve quadratic programming problem (i.e., the objective function is quadratic and the constraints are linear). He assumes a simple first order model

$$E(y) = \alpha_0, +\alpha_1 z_1 + \cdots + \alpha_p z_p \tag{12.11}$$

where $0 \leq z_i \leq 1,\ i = 1, \ldots, p$, and $\sum_{i=1}^{p} z_i = 1$. A "Target is best" scenario is assumed, so Berliner calls this problem a "control" problem, although as it will be evident, this is an optimization and not an on-line control problem in the sense described in Chapter 1. We let y denote deviations from some target T and rewrite (12.11) as follows:

$$E(y) = \alpha_0 \underbrace{\sum_{i=1}^{p} z_i}_{=1} + \sum_{i=1}^{p} \alpha_i z_i = \sum_{i=1}^{p} (\alpha_0 + \alpha_i) z_i \equiv \sum_{i=1}^{p} \theta_i z_i \tag{12.12}$$

so that $\theta_i = \alpha_0 + \alpha_i$. We assume a preliminary model like (12.12) has been fit. The model in vector form, for all n observations can then be written as

$$\boldsymbol{y} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_p)$$

so that the least squares estimator is

$$\widehat{\boldsymbol{\theta}}_{LS} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

and as we known, $\widehat{\boldsymbol{\theta}}_{LS} \sim N(\boldsymbol{\theta}, \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1})$, a density function we denote as $f(\widehat{\boldsymbol{\theta}}_{LS}|\boldsymbol{\theta}, \sigma^2)$.

The optimization problem posed by Berliner is to minimize the square of a future observation, $y = \boldsymbol{\theta}'\boldsymbol{x} + e$, where $\boldsymbol{x} = (x_1, \ldots, x_p)$ is such that $0 \leq x_i \leq 1,\ i = 1, \ldots, p$, $\sum_{i=1}^{p} x_i = 1$ and $e \sim (0, \eta^2)$, the error of the future observation, is assumed independent of $\widehat{\boldsymbol{\theta}}$. The expected loss function is then

$$E(y^2) = E_{\boldsymbol{\theta},e}(\boldsymbol{\theta}'\boldsymbol{x} + e)^2 = E(\boldsymbol{\theta}'\boldsymbol{x})^2 + \eta^2$$

so that the loss is $L(\boldsymbol{\theta}, \boldsymbol{x}) = (\boldsymbol{\theta}'\boldsymbol{x})^2$.

At this point, let us introduce a prior for the parameters, $\pi(\boldsymbol{\theta}, \sigma)$. The posterior of the parameters is then

$$\pi(\boldsymbol{\theta}, \sigma^2|\boldsymbol{\theta}_{LS}) = \frac{f(\widehat{\boldsymbol{\theta}}_{LS}|\boldsymbol{\theta}, \sigma^2)\pi(\boldsymbol{\theta}, \sigma^2)}{\int f(\widehat{\boldsymbol{\theta}}_{LS}|\boldsymbol{\theta}, \sigma^2)\pi(\boldsymbol{\theta}, \sigma^2) d\boldsymbol{\theta} d\sigma^2}.$$

Note how the dependency on the data of the posterior is through the least squares parameter estimates. The marginal posterior of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{LS}) = \int \pi(\boldsymbol{\theta}, \sigma^2|\widehat{\boldsymbol{\theta}}_{LS}) d\sigma^2$$

thus $\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{LS} \sim (\boldsymbol{\mu}(\boldsymbol{\theta}) = \widehat{\boldsymbol{\theta}}_{LS}, \boldsymbol{C}(\boldsymbol{\theta}))$ where $\boldsymbol{C}(\boldsymbol{\theta})$ is the posterior covariance matrix of $\boldsymbol{\theta}$. If σ^2 is known, then $\boldsymbol{C}(\boldsymbol{\theta})) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

The problem can now be stated in terms of the posterior expected loss:

$$\min \quad \mathcal{L}(\widehat{\boldsymbol{\theta}}_{LS}) = \int (\boldsymbol{\theta}'\boldsymbol{x})^2 \pi(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}_{LS}) d\boldsymbol{\theta}$$

subject to:

$$x_i \geq 0, \qquad \boldsymbol{x}'\mathbf{1} = 1$$

where $\mathbf{1}$ denotes a vector of ones. Since $(\boldsymbol{\theta}'\boldsymbol{x})^2 = \boldsymbol{x}'(\boldsymbol{\theta}\boldsymbol{\theta}')\boldsymbol{x}$, we have that

$$\mathfrak{l}(\widehat{\boldsymbol{\theta}}_{LS}) = \boldsymbol{x}'\boldsymbol{B}(\widehat{\boldsymbol{\theta}}_{LS})\boldsymbol{x}$$

where $\boldsymbol{B}(\widehat{\boldsymbol{\theta}}_{LS}) = \boldsymbol{C}(\widehat{\boldsymbol{\theta}}_{LS}) + \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}_{LS})\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}_{LS})'$.

Therefore, we can finally write the problem we wish to solve as (omitting the dependencies on the least squares estimates):

$$\min \quad \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x} \tag{12.13}$$

subject to

$$x_i \geq 0, \qquad \boldsymbol{x}'\mathbf{1} = 1.$$

As noted by Berliner, this is is a quadratic programming problem which has a unique solution (global minimum) if $\boldsymbol{B}$ is positive definite.

Example. Gasoline blending. Berliner [8] illustrates his approach with an example on gasoline blending. Here the response y is the octane level of the blend as a function of five different components. The variance of the octane is assumed known and to equal 0.25. The desired octane target value ranged in the (97,101) interval, so we will use a target of $T = 99$ in what follows. The experimental data is shown in Table 12.10.

We first compute the OLS estimates:

$$\widehat{\boldsymbol{\theta}}_{LS} = \begin{pmatrix} 1.45 \\ -0.85 \\ 8.88 \\ -4.24 \\ 2.72 \end{pmatrix}$$

Table 12.10. Gasoline blending Data used by Berliner [8]. To obtain y, subtract T (99) to the "response" column

z_1	z_2	z_3	z_4	z_5	response
0.000	0.000	0.350	0.600	0.060	100.0
0.000	0.300	0.100	0.000	0.600	101.0
0.000	0.300	0.000	0.100	0.600	100.0
0.150	0.150	0.100	0.600	0.000	97.3
0.150	0.000	0.150	0.600	0.100	97.8
0.000	0.300	0.049	0.600	0.051	96.7
0.000	0.300	0.000	0.489	0.211	97.0
0.150	0.127	0.023	0.600	0.100	97.3
0.150	0.000	0.311	0.539	0.000	99.7
0.000	0.300	0.285	0.415	0.000	99.8
0.000	0.080	0.350	0.570	0.000	100.0
0.150	0.150	0.266	0.434	0.000	99.5
0.150	0.150	0.082	0.018	0.600	101.9
0.000	0.158	0.142	0.100	0.600	100.7
0.000	0.000	0.300	0.461	0.239	100.9
0.150	0.034	0.116	0.100	0.600	101.2
0.068	0.121	0.175	0.444	0.192	98.7
0.067	0.098	0.234	0.332	0.270	100.5
0.000	0.300	0.192	0.208	0.300	100.2
0.150	0.150	0.174	0.226	0.300	100.6
0.075	0.225	0.276	0.424	0.000	99.1
0.075	0.225	0.000	0.100	0.600	100.4
0.000	0.126	0.174	0.600	0.100	98.4
0.075	0.000	0.225	0.600	0.100	98.2
0.150	0.150	0.000	0.324	0.376	99.4
0.000	0.300	0.192	0.508	0.000	98.6

and

$$(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \begin{pmatrix} 9.41 & 1.78 & 0.66 & -1.62 & -1.48 \\ 1.78 & 3.23 & 0.48 & -1.03 & -1.15 \\ 0.66 & 0.48 & 3.38 & -1.39 & -0.44 \\ -1.62 & -1.03 & -1.39 & 1.11 & 0.46 \\ -1.48 & -1.15 & -0.44 & 0.46 & 0.92 \end{pmatrix}.$$

The $\boldsymbol{B}(\widehat{\boldsymbol{\theta}}_{LS})$ matrix is then given by

$$\begin{aligned} \boldsymbol{B}(\widehat{\boldsymbol{\theta}}_{LS}) &= \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1} + \widehat{\boldsymbol{\theta}}_{LS}\widehat{\boldsymbol{\theta}}'_{LS} \\ &= \begin{pmatrix} 2.68 & -1.11 & 12.90 & -6.24 & 3.85 \\ -1.11 & 0.92 & -7.49 & 3.53 & -2.38 \\ 12.90 & -7.49 & 79.13 & -37.78 & 24.18 \\ -6.24 & 3.53 & -37.78 & 18.07 & -11.53 \\ 3.85 & -2.38 & 24.18 & -11.53 & 7.48 \end{pmatrix} \end{aligned}$$

which is positive definite, and hence, a unique solution exists. Solving the optimization problem (12.13), we obtain the solution[15]

$$\boldsymbol{x}^* = \begin{pmatrix} 0.0660 \\ 0.1597 \\ 0.1553 \\ 0.4343 \\ 0.1847 \end{pmatrix}$$

with an expected loss value of 0.00263. The solution turns out to be insensitive to considerable variations in the prior estimate of σ (see Problem (5)). ■

12.9 Model-Robust Bayesian Experimental Design

DuMouchel and Jones [50] proposed a simple Bayesian D-optimal criterion for experimental designs that protects against *potential* terms that one suspect may exist in the model. The model is then composed of two classes of terms: q potential terms and p *primary* terms we are more certain of their significance. The coefficients of the primary terms are assumed to have a diffuse (improper) prior with zero mean and infinite variance since no specific direction of their

[15] Any gradient-based optimization method will suffice as convergence to the global minimum is guaranteed in this example.

effect is assumed. The coefficients of potential terms, in contrast, are assumed a proper distribution with zero mean and a *finite* variance of $\tau^2\sigma^2$ that limits their actual effect. Here, τ is a tuning constant selected by the user/experimenter. As one increases τ, the variance of the prior distribution $\tau^2\sigma^2$ increases, and this implies one is expecting the potential terms to have increasingly larger effects and therefore be part of the "true" model we expect to correspond to the the actual process under study.

In this formulation, the Bayesian parameter estimates (obtained from the posterior distribution of the parameters) are given by:

$$\widehat{\boldsymbol{\beta}}_{bayes} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K}/\tau^2)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

where $\boldsymbol{K}$ is a square matrix (of dimensions equal to those of $\boldsymbol{X}'\boldsymbol{X}$, that is, $(p+q)\times(p+q)$) that has zeroes everywhere except at the q diagonal entries that correspond to the potential terms, where $\boldsymbol{K}$ contains ones. Since the covariance matrix of the parameter estimates is

$$\sigma^2(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K}/\tau^2)^{-1},$$

DuMouchel and Jones [50] suggest to find experimental designs that maximize the determinant

$$|\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K}/\tau^2|.$$

If τ^2 is small, then the prior is quite concentrated around zero, and the designs that will result will be close to the usual D-optimal designs for a model that contains the more certain terms (see Chapter 5). As τ^2 increases, the resulting Bayesian D-optimal designs will provide designs that are increasingly more appropriate to fit a model that also contains the potential terms. A "default" value of $\tau = 1$ was suggested in the absence of any other prior information [50].

12.10 Computer Implementation of Some Bayesian Optimization Methods

In this section we briefly describe some of the programs used in this chapter. They are all available at the author's personal web page.

Matlab Program for the Bayesian Optimization of a Single Response Process

The MATLAB program `maxArea` maximizes the probability that a response lies between two bounds, `ylow` and `yhigh`. The vector of observed responses, `yvec` and the `Xd` matrix containing all the corresponding regressors (in model form) need to be specified. The program call is

```
maxArea(Xd,yvec,ylow,yhigh)
```

The program requires Matlab's Statistics and Optimization toolboxes. It calls the `area` function, which computes the area under a noncentral t distribution, and the `noncentralT` function, which evaluates the noncentral T density function.

Matlab Program for the Implementation of Gilmour and Mead's [61] Stopping Criterion

The Matlab program `ComputeGainConjugate` implements Gilmour and Meads [61] approach for the computation of the expected gain after conducting a series of designed experiments. It assumes a quadratic polynomial, a maximization problem, and a conjugate prior on the parameters of the model. The program returns an array with the posterior of $L(\widehat{\boldsymbol{x}_{max}})$, for all simulated cases (`LAll`), and for those cases which resulted from concave $Y(\boldsymbol{x})$ functions, discarding the non-concave ones (`LNeg`). The function requires the $\boldsymbol{X}$ matrix with the columns in the following order: k main effect columns, $k(k-1)/2$ 2-factor interaction columns in the order $x_1x_2, x_1x_3, \ldots, x_1x_k; x_2x_3, x_2x_4, \ldots, x_2x_k; \ldots x_{k-1}x_k$; and k columns for the pure quadratic terms. This matrix and the $\boldsymbol{y}$ vector should be zeroes if one wishes to compute the expected gains before running any experiment, in which case the program will use only the prior distribution. The function call is:

```
[LAll,LNeg]=ComputeGainConjugate(X,y,Beta0,A0,v0,
sigma02,noOfSimulations,k,N)
```

where `Beta0`, `A0`, `v0`, and `sigma02` define the prior distribution (see Section 12.1.2) and the rest of the function parameters are self-explicatory. This function requires Matlab's Statistics toolbox.

Matlab Program for the Integration of the Bayesian Multivariate Predictive Density

The Matlab program `SimulateMultiPred` computes the posterior probability that p responses fall within two bounds given by vectors `low` and `high` at a given point `x`. It assumes a non-informative prior and a quadratic response model for all responses. The function call is

```
[Ytilde,prob]=SimulateMultiPred
              (X,Y,x,noOfSimulations,low,high);
```

The matrix `X` has the experimental design in model form (a column of ones, all main effects, all 2-factor interactions and all pure quadratic columns. The matrix `Y` contains is the response values for all the responses, one per column. The point `x` is not in model form, it only has the coordinates of all the controllable factors. The program returns `prob`, the posterior probability of conformance, and `Ytilde`, a vector of predictive densities at `x` obtained from simulating (specify `noOfSimulations`). The program requires Matlab's Statistics toolbox.

This program can be easily extended into a program that finds the point $\boldsymbol{x}$ at which the posterior probability of the responses falling between their bounds is maximized.

12.11 Problems

1 Derive the marginal posterior density of $\boldsymbol{\beta}$, the vector of parameters in a linear regression model. That is, derive the density of $\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}$. Assume non-informative priors.

2 Consider a 3-factor CCD run with one center point and $\alpha = 2$. The factorial part is a 2^3 and a full quadratic model is going to be fit. The 15 observations of the response are: 45.9, 60.6, 57.5, 58.6, 53.3, 58, 58.8, 52.4, 46.9, 55.4, 55, 57.5, 56.3, 58.9, and 56.9. These correspond to the factorial runs (in standard order), followed by the $-\alpha, +\alpha$ axial runs for each factor, followed by the center point. Using non-informative priors, find the posterior predictive density of the next response value assuming non-informative priors. Give a 95% credible interval on this density at the point $\boldsymbol{x}' = (-1.4, 2.6, 0.7)$.

3 Consider the following data for the sales of a product (y) as a function of R&D spending (x), both in million of dollars:

y	x
25	0.10
25	0.15
25	0.20
50	10.00
50	20.00
50	32.00
50	40.00
100	50.00
100	60.00
100	64.00
100	64.00
120	75.00
200	75.00
300	100.00
350	128.00
350	128.00
400	150.00
500	256.00
700	300.00
700	300.00
1000	400.00

If a simple first order model $y = \beta_0 + \beta_1 x + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ is going to be fit to these data, find a 95% credible interval for the next sales value which we could predict if $x = 500$ is spent in R&D. Use non-informative priors in all parameters $(\beta_0, \beta_1, \sigma)$.

Solve:

a) analytically, i.e., from the closed-form expression of the predictive density;

b) from simulating the predictive density by simulating, in turn, the posterior of the parameters and the likelihood function.

In each case, show a plot of the predictive density and indicate the limits of the credible interval.

4 Consider the metal cutting example in Chapter 1 and the `ComputeGain-Conjugate` Matlab program. Use this program to find the expected gains in that optimization exercise using Gilmour and Mead's method:

a) before running any experiment;

b) after running the factorial points;

c) after running the axial points but before running the center points;

d) after running the center points.

In all cases, assume the following prior: $\boldsymbol{\beta}_0' = (10, -5, -5, -5, 0, 0, 0, 0, 0, 0)$, $\sigma_0^2 = 5$ and $\boldsymbol{A}_0 = \boldsymbol{I}$. Solve each case for increasingly flatter priors.

5 Solve the gasoline blending mixture optimization problem of Section 12.8 for: $\sigma = 0.40$ and $T = 99$, $\sigma = 0.25$ and $T = 101$, and $\sigma = 0.40$ and $T = 101$.

PART VI

INTRODUCTION TO OPTIMIZATION OF SIMULATION AND COMPUTER MODELS

Chapter 13

SIMULATION OPTIMIZATION

Experiments are mediators between nature and ideas.

—Johan Wolfgang Goethe (1749–1832)

13.1 Introduction

This book thus far has focused on the optimization of industrial processes where a physical system or process exists and needs to be improved. There is a growing awareness in the Statistics and Engineering literature for the need and usefulness of methods for the optimization of *models* of such physical processes. The models are a surrogate of a process or product, used to study and improve it with no active intervention and faster experimentation. When optimizing a simulation model, the optimal solution obtained from the simulation is implemented in the real system. Evidently, the model must be an accurate representation of the system under study. We will not delve into the deep subject of simulation modeling and validation, for which a very large body of literature exists (see e.g., the books [83, 4, 143]) and is outside the scope of the present book. Our purpose in this chapter is to provide an introduction to some of the techniques that are useful in the optimization of simulated systems.

One peculiarity of a simulated process is that it does not need so much human attention as a real or physical process. It has been emphasized by Response Surface authors that RSM methods must be carefully used and should not be thought of as a computer algorithm that is blindly applied. The process of knowledge discovery entails a continuous interplay between planning and

conducting an experiment and analyzing the results of the experiment, planning further experiments and so on [25]. This process also applies to a simulated system, and traditional RSM techniques have been widely used in simulation optimization. However, given that we work with a computer model rather than with a physical system, the optimization of a simulated system can be done in a more automatic way, and more formal algorithmic methods are also applicable.

In this chapter, we first introduce optimization methods based on stochastic approximation techniques. Given their algorithmic form, these methods are sometimes appropriate for simulation optimization. Contrary to RSM methods, which are much less mathematically structured, the kind of stochastic optimization methods described here are known to converge to a stationary point under certain conditions, in a similar way that nonlinear programming techniques are known to converge for the optimization of deterministic functions. The last section discusses the *statistical* test of the Karush-Khun-Tucker (KKT) necessary optimality conditions, which provides a stopping criterion in constrained simulation optimization problems.

13.2 Newton's Method and Stochastic Approximation

Consider first Newton's method for solving for the root of a function, i.e., find x (a scalar) such that $f(x) = 0$ is true. Here $f(x)$ is a deterministic function from $\Re$ to $\Re$, in other words, $f(x)$ can be measured without error. The recursive equation that is known to converge to the root is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{13.1}$$

where $f'(x) = df(x)/dx$ is known analytically. In this setting, convergence simply means that $x_n \to \theta$ as $n \to \infty$ (where θ is the closest root to x_1). The recursion is equivalent to steepest descent applied to $f(x)$.

The seminal work of Robbins and Monro (RM) [135] extended such approach to a stochastic function, that is, assume $y = y(x) = M(x) + \varepsilon$ where ε is a random variable with $E[\varepsilon] = 0$ so $M(x) = E[y|x]$ is the regression of y on x, a function which can be nonlinear in x. No parametric assumptions are made on the form of $M(x)$. Suppose we want to find a root of the equation $M(x) = 0$. Neither $M(x)$ nor $M'(x)$ are known analytically, and we cannot observe $M(x)$ directly, we can only observe $y(x)$. Suppose that instead of

using (13.1) (which is not possible to use since the functions needed are not available to us) we use the recursion (sometimes called the "RM process"):

$$x_{n+1} = x_n - a_n y(x_n) \tag{13.2}$$

where the sequence $\{a_n\}_{n=1}^{\infty}$ needs to be determined and x_1 is selected arbitrarily. This recursion determines an experimental design on the factor space (a one dimensional space when x is a scalar) that will seek the root of the equation. Robbins and Monro asked a very simple question: under what conditions on $M(x)$, ε, and $\{a_n\}_{n=1}^{\infty}$ do we have convergence of the sequence of experimental factor levels $\{x_n\}_{n=1}^{\infty}$ to the root of $M(x) = 0$, let us call it θ? In their original paper, they showed that, if

a)

$$a_n \to 0, \quad \sum_{n=1}^{\infty} a_n = \infty$$

b)

$$\sum_{n=1}^{\infty} a_n^2 < \infty$$

c) the regression function is such that $M(x) \leq 0$ if $x < \theta$ and $M(x) \geq 0$ if $x > \theta$,

d) the distribution function of ε has zero mean and bounded tails,

then using (13.2) we have

$$\lim_{n \to \infty} E[(x_n - \theta)^2] = 0$$

i.e., they show convergence in *mean square* of x_n to the root θ. A sequence that satisfies the first two conditions is the harmonic series $a_n = 1/n = \{1, 1/2, 1/3, \ldots\}$. Condition a) is needed because otherwise the search could stop before finding a root. Condition b) is needed to eventually eliminate the "noise" in the observations. Some of the assumptions under which mean square convergence is achieved were simplified and clarified by Dvoretzky [51]. In particular, instead of assumption d) it is only necessary that $E[\varepsilon] = 0$ and $\sigma_\varepsilon^2(x) < \infty$. Note that the variance can be non-homogeneous, a condition of considerable interest in response surface methods. The Dvoretzky conditions imply both MS convergence and convergence of x_n to θ with probability one.

Note how these conditions are *sufficient*, and therefore, several authors have come up with slightly different sufficient conditions (see [145] for a recent account).

13.3 Stochastic Gradient

Following a suggestion by Robbins and Monro, Kiefer and Wolfowitz [79] proposed to apply Robbins and Monro's stochastic approximation approach to finding the root of $M'(x) = 0$, i.e., for finding the stationary point θ of the regression function that is known to have a maximum (or a minimum, as desired). Since the slope of the regression function is not observable directly, they propose to use instead:

$$x_{n+1} = x_n + a_n \frac{y(x_n + c_n) - y(x_n - c_n)}{c_n} \tag{13.3}$$

(if we wish to minimize we subtract the second term instead). Kiefer and Wolfowitz showed that x_n as above converges in mean square to θ if

$$a_n \to 0, \quad c_n \to 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \qquad \sum_{n=1}^{\infty} a_n c_n < \infty, \qquad \sum_{n=1}^{\infty} a_n^2 c_n^2 < \infty \tag{13.4}$$

besides of some regularity conditions on $M(x)$.

Blum [12] extended this procedure to the multivariate case, of particular interest in realistic process optimization problems.

We now present the algorithmic details of a multivariate generalization of Kiefer and Wolfowitz' stochastic gradient, following Spall [145], who refers to the following method as the *Finite Difference Stochastic Approximation* (SA) method.

Suppose we observe $y(\boldsymbol{x}) = L(\boldsymbol{x}) + \varepsilon(\boldsymbol{x})$, where $\boldsymbol{x}$ is a $k \times 1$ vector of controllable factors and $L(\boldsymbol{x})$ is some loss function we wish to minimize, without loss of generality. Thus, we cannot measure the loss directly, we only have noisy measurements where $\varepsilon(\boldsymbol{x})$ is the process noise. The multivariate recursion, analogous to (13.3), is given by

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - a_n \widehat{\boldsymbol{g}}_n(\boldsymbol{x}_n) \tag{13.5}$$

where the gradient estimate is given by

$$\widehat{\boldsymbol{g}}_n(\boldsymbol{x}_n) = \begin{bmatrix} \frac{y(\mathbf{x}_n + c_n \mathbf{1}_{(1)}) - y(\mathbf{x}_n - c_n \mathbf{1}_{(1)})}{2c_n} \\ \vdots \\ \frac{y(\mathbf{x}_n + c_n \mathbf{1}_{(k)}) - y(\mathbf{x}_n - c_n \mathbf{1}_{(k)})}{2c_n} \end{bmatrix} \tag{13.6}$$

and where $\mathbf{1}_{(i)}$ is a $k \times 1$ vector with a one in position i and zeroes elsewhere. Spall [145] suggests the following scheme for the "gains", a_n and c_n:

$$a_n = \frac{a}{(n+1+A)^{\alpha}}, \quad c_n = \frac{c}{(k+1)^{\gamma}}$$

and suggests the values $0.6 < \alpha \leq 1$, $0.1 < \gamma < 0.5$, $\alpha - \gamma > 0.5$ (choices suggested by Spall that meet these conditions are $\alpha = 0.602$ and $\gamma = 0.101$). The value of c is suggested to be around the standard deviation of the noise term (σ_ε) and the value of A is suggested to be about 10% the number of iterations one wishes to apply the procedure. The key recommendation for c has the merit of scaling the procedure and to make it insensitive to the noise. If the noise variance is unknown, and the value of c is far from the actual σ_ε^2, the performance will be much different (see example below). Spall also suggests to calibrate the parameter a such that the initial change of the controllable factors, as given by $a\widehat{\boldsymbol{g}}_0(\boldsymbol{x}_0)/(1+A)^{\alpha}$, is as desired, although no recommendation about how to choose this desirable value is given, other than suggesting it should be small. Evidently, larger values of a will increase the "step size".

Example. Maximization of a quadratic function. To illustrate the multivariate stochastic gradient method, let us consider the function $L(\boldsymbol{x}) = 773.8 - 7.2x_1 + 8.2x_2 + 166.5x_1^2 + 189.9x_2^2$ which is observed with additive error $\varepsilon \sim N(0, 10^2)$. The factors are coded using $x_1 = (\xi_1 - 275)/50$ and $x_2 = (\xi_2 - 158.9)/50$. The function $L(\boldsymbol{x})$ is negative definite, and the maximum is at $\boldsymbol{\xi}' = (276.08, 157.82)$, in original units. The parameters $\alpha = 0.602, \gamma = 0.101$, $a = 0.9$ were used for 500 iterations, thus $A = 50$. The initial point $\boldsymbol{\xi}_0' = (200, 100)$ was used with the "'perfect" value of $c = \sigma_\epsilon = 10$. Figure 13.1 shows 10 realizations of the search process. The search stops slightly earlier than the exact maximum. Using 1000 iterations and keeping all other parameters equal reaches the maximum. Suppose instead that σ_ϵ is unknown and we use instead $c = 1 < \sigma_\varepsilon = 10$. The same figure shows 10 realizations of the SA process if started from the point $\boldsymbol{\xi}_0' = (350, 100)$. As it can be seen, the performance is much more erratic. ■

13.4 Stochastic Perturbation Stochastic Approximation

The stochastic gradient method approximates the gradient of the function by running the process at $2k$ points forming a "star". Spall [144] noted that the number of function evaluations, or simulations in simulation-optimization

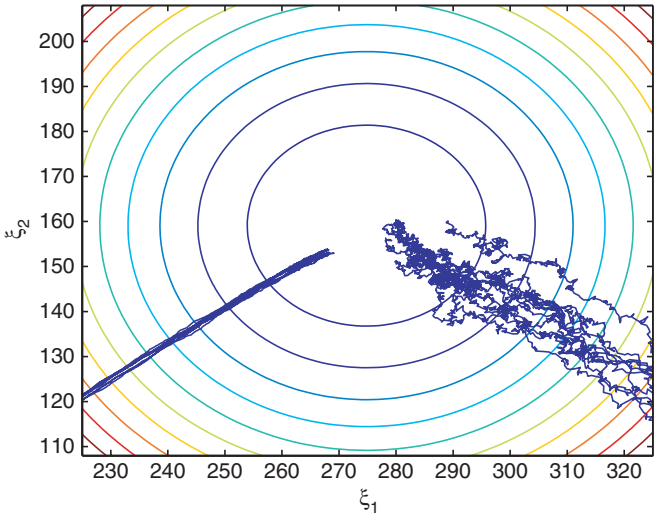


Figure 13.1. Trajectories of the stochastic gradient search, 500 iterations in each trajectory. The 10 trajectories on the left started from $\boldsymbol{\xi}_0' = (200, 100)$ with $c = \sigma_\epsilon$; the 10 trajectories on the right started from $\boldsymbol{\xi}_0' = (350, 100)$ with $c = 1 < \sigma_\epsilon$ which makes the method very sensitive to the large errors

applications, can be reduced by a factor of k if only *two* measurements are taken at each iteration. The method, called Stochastic Perturbation Stochastic Approximation (SPSA), uses the same basic iteration (13.5) while estimating the gradient using instead

$$\widehat{\boldsymbol{g}}_n(\boldsymbol{x}_n) = \begin{bmatrix} \frac{y(\mathbf{x}_n + c_n\boldsymbol{\Delta}_n) - y(\mathbf{x}_n - c_n\boldsymbol{\Delta}_n)}{2c_n\boldsymbol{\Delta}_{n1}} \\ \vdots \\ \frac{y(\mathbf{x}_n + c_n\boldsymbol{\Delta}_n) - y(\mathbf{x}_n - c_n\boldsymbol{\Delta}_n)}{2c_n\boldsymbol{\Delta}_{nk}} \end{bmatrix} \tag{13.7}$$

where $\boldsymbol{\Delta}_n$ is a $k \times 1$ random perturbation with certain properties that allow for convergence. The easiest way to satisfy such conditions, and it seems the most popular choice of distribution for $\boldsymbol{\Delta}$, is a symmetric Bernoulli distribution of the form

$$\Delta_i = \begin{cases} -1 & \text{with prob. } = 0.5 \\ 1 & \text{with prob. } = 0.5 \end{cases}$$

for $i = 1, 2, \ldots, k$.

Since the numerator in each entry of the $\widehat{\boldsymbol{g}}_n(\boldsymbol{x}_n)$ vector is the same, *only two function evaluations are necessary per iteration*. This makes the

method specially attractive for cases where each function evaluation (i.e., each simulation) is very expensive or too time consuming to conduct. The ratio of the number of function evaluations in the SPSA algorithm to that of the SA algorithm is therefore $1/k$.

Spall [145] suggests using the same scheme for the a_n and c_n gains as in stochastic gradient. Thus, if the standard deviation is known, setting $c = \sigma_\epsilon$ scales the algorithm and makes it very robust against noise.

The key insight behind the SPSA algorithm is that average gradient information can be built *across* "experiments" and that it suffices to use only two observations per experiment or iteration of the method.

Example. Maximization of a quadratic function. To illustrate the SPSA algorithm, consider the same example we used to demonstrate the SA algorithm. Figure 13.2 shows trajectories similar than those in Figure 13.1 under similar conditions. As it can be seen, the behavior is very similar, although SPSA utilizes exactly *half* the number of function evaluations (500/2, since $k = 2$). ■

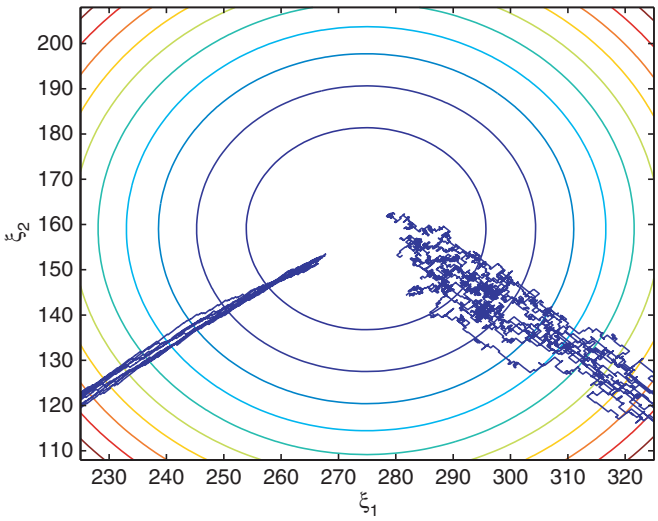


Figure 13.2. Trajectories of the SPSA search, 500 iterations in each trajectory. The 10 trajectories on the left started from $\boldsymbol{\xi}_0' = (200, 100)$ with $c = \sigma_\epsilon$; the 10 trajectories on the right started from $\boldsymbol{\xi}_0' = (350, 100)$ with $c = 1 < \sigma_\epsilon$. In either case the behavior is approximately similar to that of the SA search, but SPSA uses half (250) function evaluations

13.5 Statistical Test of KKT Optimality Conditions in Simulation Optimization**

Consider[1] a stochastic (perhaps, discrete event) simulation of a system e.g., a simulation model of a manufacturing process. Such system can have multiple responses of interest, probably correlated, of unknown form. Suppose the ultimate goal is to solve, via simulation-optimization the following problem:

$$\min E[Y_0(\boldsymbol{x}, r)] \tag{13.8}$$

$$\text{subject to:} \quad E[Y_h(\boldsymbol{x}, r)] \geq a_h, \quad h = 1, 2, \ldots, m-1, \tag{13.9}$$

where $\boldsymbol{x}$ is a $k \times 1$ vector of controllable factors and r is a pseudo-random number sequence used in the simulation program that gives randomness to the functions. This problem is also relevant in Response Surface Methodology where physical experiments are conducted, but we will confine ourselves to the simulation optimization scenario in this section.

The way to solve problem (13.8–13.9) using a simulation code is to perform experiments (simulations) at different settings $\boldsymbol{x}$ and hope to find a point which satisfies the Karush-Kuhn-Tucker (KKT) necessary conditions for optimality (see Appendix C):

$$\boldsymbol{\beta}_{-0;0} = \boldsymbol{B}_{-0;J}\boldsymbol{\lambda} \qquad \boldsymbol{\lambda} \geq \boldsymbol{0}$$

where $\boldsymbol{\beta}_{-0;0}$ is the gradient of the goal function, $\boldsymbol{B}_{-0;J}$ is the $k \times J$ matrix with the gradients of the J binding (active) constraints, and λ is a $J \times 1$ vector containing the Lagrange multipliers. Recall that the KKT conditions are *necessary*, that is, if $\boldsymbol{x}$ optimal, $\boldsymbol{x}$ satisfies KKT. This also implies that if the KKT conditions do not hold at $\boldsymbol{x}$, then $\boldsymbol{x}$ is not optimal (so we need to keep searching). The question is how to test these conditions from noisy experimental data.

In a recent paper, Bentovil et al. [9] proposed an effective method to test the KKT conditions in a simulation optimization problem for problem (13.8–13.9). It is assumed that a region in the space of controllable factors have been reached (perhaps by a line search technique such as steepest ascent) where second order models are adequate for all response functions in the problem. The procedure is as follows:

1 Run the simulation using DOE around the point of interest that allows to fit quadratic models to all the responses.

[1]This section contains somewhat more advanced material and may be skipped on a first reading.

2 Perform a statistical test for determining all binding constraints (using a t-test). This gives J. Perform also a lack of fit test.

3 If all models fit well, then test the KKT conditions. The test is:

$$H_0 : E(\widehat{\boldsymbol{\beta}}_{-0;0} - \widehat{\boldsymbol{\mathcal{B}}}_{-0;J}\widehat{\boldsymbol{\lambda}}) = \mathbf{0}$$

vs. a two sided alternative. Since the statistic $\widehat{\boldsymbol{\beta}}_{-0;0} - \widehat{\boldsymbol{\mathcal{B}}}_{-0;J}\widehat{\boldsymbol{\lambda}}$ is quite non-linear and does not have a known distribution, we approximate its distribution via *Parametric Bootstrapping* [53].

The parametric bootstrapping approach is used to test the hypothesis as follows:

1 From the fitted models and the covariance matrix of the parameter estimates of the objective function and active constraints, simulate (i.e., resample) R instances of the bootstrapped parameters, namely $\widehat{\boldsymbol{\beta}}^*_{-0;0}$, $\widehat{\boldsymbol{\mathcal{B}}}^*_{-0;J}$ and compute

$$\widehat{\boldsymbol{\lambda}^*} = (\widehat{\boldsymbol{\mathcal{B}}}^{*'}_{-0;J}\widehat{\boldsymbol{\mathcal{B}}}^*_{-0;J})^{-1}\widehat{\boldsymbol{\mathcal{B}}}^{*'}_{-0;J}\widehat{\boldsymbol{\beta}}^*_{-0;0}.$$

2 With the R bootstrapped values compute the errors $\boldsymbol{e} = \widehat{\boldsymbol{\beta}}^*_{-0;0} - \widehat{\boldsymbol{\mathcal{B}}}^*_{-0;J}\widehat{\boldsymbol{\lambda}^*}$ and form the multivariate Empirical Distribution Function (EDF) of the errors.

3 If $\mathbf{0}$ is not inside the $100(1-\alpha)\%$ bootstrapped confidence region for

$$\boldsymbol{\eta} = E[\widehat{\boldsymbol{\beta}}_{-0;0} - \widehat{\boldsymbol{\mathcal{B}}}_{-0;J}\widehat{\boldsymbol{\lambda}}],$$

formed by the $\alpha/2$-percentile and the $1-\alpha/2$-percentiles of the EDF of the errors, then reject H_0 at the current point (i.e., KKT conditions do not hold).[2]

We need to test also that $\boldsymbol{\lambda} \geq 0$ to complete the KKT conditions. This can be done with a binomial test from the bootstrapped $\widehat{\boldsymbol{\lambda}^*}$'s (see [9]).

The rationale for this procedure is the geometrical meaning of the KKT conditions (see Appendix C). The conditions state that at the point of interest the

[2]One conservative but easy way to do this is to test that each element e_i of $\boldsymbol{e}$ falls inside the individual confidence interval given by the $(\alpha/2)/k$ and $1-(\alpha/2)/k$ percentiles of the ith coordinate of the EDF. This is called a Bonferroni approach for multiple confidence regions and is conservative since the overall confidence level will be greater or equal than $100(1-\alpha\%)$.

gradient of the objective function is a linear combination of the gradients of the constraints that are active. If zero is inside the confidence region of the errors $\boldsymbol{\eta}$, it means that such linear combination holds statistically (i.e., with $100(1-\alpha)\%$ confidence) at the current point of interest.

Example. Consider a "synthetic" problem presented in [9]. We wish to solve the following problem:

$$\text{Min} \quad E[(d_1 - 8)^2 + (d_2 + 8)^2 + \varepsilon_0]$$

subject to:

$$E[(d_1 - 3)^2 + d_2^2 + d_1 d_2 + \varepsilon_1] \leq 4$$
$$E[(d_1^2 + 3(d_2 + 1.061)^2 + \varepsilon_2] \leq 9$$

where the errors are multivariate normal with $\text{Var}(\varepsilon_0) = 1$, $\text{Var}(\varepsilon_1) = 0.15^2$, $\text{Var}(\varepsilon_2) = 0.4^2$, and $\text{Cov}(\varepsilon_0, \varepsilon_1) = 0.09$, $\text{Cov}(\varepsilon_0, \varepsilon_2) = 0.12$, and $\text{Cov}(\varepsilon_1, \varepsilon_2) = -0.006$. The optimal point is $\boldsymbol{d}^* = (2.5328, -1.9892)'$ which corresponds to point A in Figure 13.3. We test at points A, B, C, D. The test should reject the null hypothesis (which says that the KKT conditions do hold at a certain point) α % of the times at point A, and it should reject it much more often at the other points (with point B being a "better" candidate than point D, which is the worst). This gives and indication of the power of the test.

The results of the performance of the bootstrapping approach for this problem are shown in Table 13.1. A $N = 12$ run Central Composite Design was run around each of the 4 points A, B, C, and D. Two different α values were tested, 5% ($R = 2000$ was used) and 10% ($R = 1000$ was used). As it can be seen from the table, the method rejects the optimum point α % of the times, as expected. Furthermore, the farther the other 3 points are from being a KKT point, the higher rejection rate of the hypothesis. This indicates adequate performance of the bootstrapping test. ■

The performance of the bootstrapping approach depends on the "signal to noise" ratio, that is, on the amount of noise the errors ε_i contain relative to the size of the region where the experimental design is run. But it should be pointed out that no method will work well when the noise overwhelms a process; in such case the KKT conditions are not testable.

An example where the bootstrapping method was applied to an Inventory System simulation is described in [9].

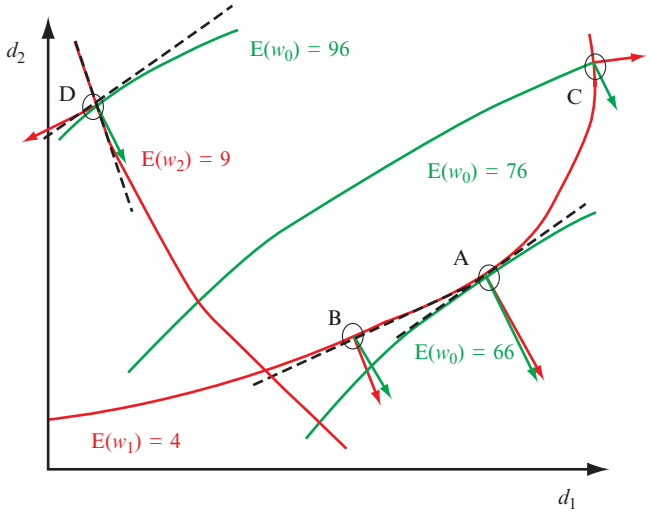


Figure 13.3. Contour graph of the example problem. Point A is optimal, while the other 3 labeled points are not. From: [9]

Table 13.1. Results of the Bootstrapping KKT Test applied to the synthetic example

	α=10%, R=1000	α=5%, R=2000
Point	% rejects	% rejects
A, (opt)	0.0942	0.0571
B	0.2208	0.1969
C	0.6509	0.5524
D	0.9976	0.9965

13.6 Problems

1 Write a Matlab program that implements the Stochastic Gradient procedure for two factors. Test your program against the examples in this chapter.

2 Using the program in Problem 1, repeat the example in this chapter by only changing the a parameter to be 5.0 (instead of 0.9). How different is the behavior if the stochastic gradient method for the example?

3 Consider the function $L(\boldsymbol{x}) = 773.8 - 7.2x_1 + 8.2x_2 + 166.5x_1^2 + 189.9x_2^2 + 20x_1x_2$ which is observed with additive error $\varepsilon \sim N(0, 10^2)$. The factors are coded using $x_1 = (\xi_1 - 275)/50$ and $x_2 = (\xi_2 - 158.9)/50$.

a) Write a Matlab program that simulates this function.

b) Use the Matlab program you developed in Problem 1 to find a stationary point using the Stochastic Gradient method. Start all runs from the point $\boldsymbol{\xi}_0' = (200, 100)$ and use $c = \sigma_\epsilon$.

c) repeat part b) starting from $\boldsymbol{\xi}_0' = (350, 100)$ with $c =< \sigma_\epsilon$

Run the simulation/optimization process in b) and c) for 500 times. Compare against the maximum of the expected response ($L(\boldsymbol{x})$.)

4 Write a Matlab program that implements the SPSA method for two factors. Test your program against the examples in this chapter.

5 Using the program in the previous problem, repeat the examples in this chapter by only changing the value of parameter a to 5.0. How different is the performance of the SPSA method for this example?

6 Consider again the function $L(\boldsymbol{x}) = 773.8 - 7.2x_1 + 8.2x_2 + 166.5x_1^2 + 189.9x_2^2 + 20x_1x_2$ which is observed with additive error $\varepsilon \sim N(0, 10^2)$. The factors are coded using $x_1 = (\xi_1 - 275)/50$ and $x_2 = (\xi_2 - 158.9)/50$.

a) Use the Matlab program you developed in Problem 3 to find a stationary point using the SPSA method. Start all runs from the point $\boldsymbol{\xi}_0' = (200, 100)$ and use $c = \sigma_\epsilon$.

b) repeat part a) starting from $\boldsymbol{\xi}_0' = (350, 100)$ with $c =< \sigma_\epsilon$

Run the simulation/optimization process in a) and b) for 500 times. Compare against the maximum of the expected response ($L(\boldsymbol{x})$.)

Chapter 14

KRIGING AND COMPUTER EXPERIMENTS

Science is nothing but the elaboration of a model for nature.

—Arturo Rosenblueth (1900–1970)

Engineering problems where complex computer codes need to be run in order to obtain a solution are increasingly common. This could be, for example, a finite-element program which evaluates certain physical properties of a mechanical design. The finite element model is accurate and deterministic in the sense that when run from the same settings it produces the exact same outputs, but may take a long time to run. Still, a faster approximation (or *metamodel*) of the computer code output is desirable, particularly for optimization purposes, since optimization would require running the code several times. The accuracy (bias) of such approximation is of prime interest. This leads to a different problem of function approximation methods, closer to what Chebyshev investigated in the XIX century. Kriging methods are one class of popular methods to provide interpolation of complex, "expensive" functions which can only be observed by running such a computer code[1]. This chapter

[1]In contrast with Chebyshev's approximations, Kriging methods seek a best approximation in the mean square error (MSE) sense. Interestingly, Chebyshev found the MSE criterion unsatisfactory for solving the type of problems in the designs of machines he was interested in, since a machine component that exceeds a tolerance, even by a small amount, would be intolerable; see [2, pp. 299–300]. Chebyshev and other approximations, such as Fourier's, seek approximation of a function in terms of combinations of other given functions; Kriging approximates a function measured at a discrete set of points with a convex combination of the observations that provides minimum MSE.

provides an introduction to Kriging methods used for prediction. We also discuss experimental design issues for computer codes, in particular, space filling designs used to reduce bias in the fitted model.

14.1 Kriging**

Kriging is a type of Spatial Statistical method useful for interpolation that originated in geosciences[2]. The method was initially developed by D.G. Krige and H.S. Sichel, a mining engineer and a statistician from South Africa, respectively, in the early 1950's, when classical statistics were found to be unsuitable for estimating ore reserves. Krige improved his methods in the 1960's in collaboration with G. Matheron, a French mathematician who coined the term Kriging and developed the methods into the 1970's and 1980's. From a process optimization perspective, Kriging can be applied for modeling response surfaces of simulation or expensive computer model output. This provides a simplified model (of the computer models), which in turn can be optimized. Here we provide an introduction to the basic Kriging methods. For a fuller treatment, see the (by now classic) book by Cressie [35].

Let $\{x_1, x_2, \ldots, x_n\}$ (or $\boldsymbol{x} \in \Re^n$) be "locations" where we observe data $\{y(x_1), y(x_2), \ldots, y(x_n)\}$. In process optimization, the variables x_i denote the controllable factors of the simulation or computer model. It is supposed the data are a realization of a stochastic process $Y(\cdot)$:

$$\{Y(\boldsymbol{x}) \ : \ \boldsymbol{x} \in D \subset \Re^k\}.$$

In *Ordinary Kriging* one assumes the process generating the data obeys the model

$$Y(\boldsymbol{x}) = \mu + \delta(\boldsymbol{x}), \quad x \in D \subset \Re^k \tag{14.1}$$

where $\mu \in \Re$ is a non-random constant and $\delta(\boldsymbol{x})$ is a spatial stochastic process with

$$E(\delta(\boldsymbol{x})) = 0 \quad \text{for all} \quad \boldsymbol{x} \in D$$

and $\text{Cov}(\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2)) = \boldsymbol{C}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in D$ (hence, $\delta(\boldsymbol{x})$ is said to be a (multivariate) white noise process). Therefore, the covariance between the errors obtained in two different locations depends only on the distance between the two points being considered through the function $C(\cdot)$, called the

[2]This section contains relatively more advanced material.

covariogram (this is analogous to the autocovariance function, see, e.g., [26], [39], and hence $\boldsymbol{C}(\mathbf{0}) = \text{Var}(Y(\boldsymbol{x}))$.) The processes $\delta(\boldsymbol{x})$ and $Y(\boldsymbol{x})$ are therefore assumed to be second order (or weakly) stationary, with $Y(\boldsymbol{x})$ such that

$$E[(Y(\boldsymbol{x})] = \mu \quad \text{for all } \boldsymbol{x} \in D$$

and

$$\text{Cov}(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2)) = C(\boldsymbol{x}_1 - \boldsymbol{x}_2).$$

Stationarity of $Y(\boldsymbol{x})$ dos not guarantee $Y(\boldsymbol{x})$ is ergodic. To be ergodic, and hence, to be able to make valid inferences based on a single "observed path" of the process, we must have $C(\boldsymbol{x}) \to 0$ as $\boldsymbol{x} \to \infty$. In most cases, this is a reasonable assumption.

An important function in Kriging is

$$\begin{aligned} \text{Var}(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2)) &= E[(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2) - E^2(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2))^2] \\ &= E[(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2))^2] \end{aligned}$$

where the last equality follows since $E(Y(\boldsymbol{x})) = \mu$ for all $\boldsymbol{x}$, thus,

$$\text{Var}(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2)) = 2\gamma(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in D$, a function called the *variogram*. The quantity $\gamma(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ is called the *semi-variogram*. Since

$$\text{Var}(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2)) = \text{Var}(Y(\boldsymbol{x}_1)) + \text{Var}(Y(\boldsymbol{x}_2)) - 2\text{Cov}(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2))$$

and since $Y(\cdot)$ is covariance stationary, then

$$\text{Var}(Y(\boldsymbol{x}_1) - Y(\boldsymbol{x}_2)) = 2[C(\mathbf{0}) - C(\boldsymbol{x}_1 - \boldsymbol{x}_2)] = 2\gamma(\boldsymbol{h})$$

where $\boldsymbol{h} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ is called the "lag" in analogy to time series analysis. Note that $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$ and that $\gamma(\mathbf{0}) = 0$.

The assumed form of the predictor at some point $\boldsymbol{x}_0 \in D$ used in Kriging is

$$\widehat{Y}(\boldsymbol{x}_0) = p(Y(\boldsymbol{x}_0)) = \sum_{i=1}^{n} \lambda_i y(\boldsymbol{x}_i), \quad \sum_{i=1}^{n} \lambda_i = 1. \tag{14.2}$$

Under these assumptions, the goal in Kriging is to find $p(Y(\boldsymbol{x}_0))$ by choosing the λ_i in (14.2) such that the prediction provides minimum mean square error (MMSE) prediction[3], i.e., it should minimize

$$E[(Y(\boldsymbol{x}_0) - p(Y(\boldsymbol{x}_0)))^2].$$

[3]Note that the optimal λ_i's will depend on the point $\boldsymbol{x}_0$ at which we are predicting.

Kriging predictions will thus constitute a "BLUE" (Best Linear Unbiased Estimator) of the underlying process $Y(\boldsymbol{x})$. Using a Lagrange multiplier m and adding a constant 2 (for convenience as we will see later), the Lagrangian is

$$L(\boldsymbol{x}_0) = E[(Y(\boldsymbol{x}_0) - p(Y(\boldsymbol{x}_0)))^2] - 2m\left(\sum_{i=1}^{n}\lambda_i - 1\right). \tag{14.3}$$

After some algebra, it is easy to show (see Exercise 13) that

$$L(\boldsymbol{x}_0) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) + 2\sum_{i=1}^{n}\lambda_i\gamma(\boldsymbol{x}_0 - \boldsymbol{x}_i) - 2m\left(\sum_{i=1}^{n}\lambda_i - 1\right). \tag{14.4}$$

Equating $\partial L(\boldsymbol{x}_0)/\partial\boldsymbol{\lambda}$ and $\partial L(\boldsymbol{x}_0)/\partial m$ to zero, we obtain a system of $n+1$ equations with $n+1$ unknowns. Define

$$\boldsymbol{\lambda}_0 = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \hline \lambda_n \\ m \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\lambda} \\ \hline m \end{pmatrix},$$

$$\boldsymbol{\gamma}_0 = \begin{pmatrix} \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_1) \\ \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_2) \\ \vdots \\ \hline \gamma(\boldsymbol{x}_0 - \boldsymbol{x}_n) \\ 1 \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\gamma} \\ \hline 1 \end{pmatrix},$$

and

$$\boldsymbol{\Gamma}_0 = \left(\begin{array}{cccc|c} \gamma(\mathbf{0}) & \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_2) & \ldots & \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_n) & 1 \\ & \gamma(\mathbf{0}) & \ldots & \gamma(\boldsymbol{x}_2 - \boldsymbol{x}_n) & 1 \\ & & \ddots & & 1 \\ \hline symm. & & & \gamma(\mathbf{0}) & 0 \end{array}\right) \equiv \left(\begin{array}{c|c} \boldsymbol{\Gamma} & \mathbf{1} \\ \hline \mathbf{1}' & 0 \end{array}\right).$$

Thus, we have that

$$\boldsymbol{\lambda}_0 = \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\gamma}_0.$$

It can be shown, after some algebra (see Exercise 14) that

$$\boldsymbol{\lambda}' = \left(\boldsymbol{\gamma} + \mathbf{1}\frac{1 - \mathbf{1}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}}{\mathbf{1}'\boldsymbol{\Gamma}^{-1}\mathbf{1}}\right)'\boldsymbol{\Gamma}^{-1} = (\boldsymbol{\gamma} - \mathbf{1}\ m)'\boldsymbol{\Gamma}^{-1}. \tag{14.5}$$

so $m = -\frac{1-\mathbf{1}'\mathbf{\Gamma}^{-1}\boldsymbol{\gamma}}{\mathbf{1}'\mathbf{\Gamma}^{-1}\mathbf{1}}$. However, the semi-variogram $\gamma(\boldsymbol{h})$ is unknown and needs to be estimated from data. A common way of doing this is to use a method of moments estimator:

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|}\sum_{N(\boldsymbol{h})}(Y(\boldsymbol{x}_i) - Y(\boldsymbol{x}_j))^2, \quad \boldsymbol{h} \in \Re^n \tag{14.6}$$

where $N(\boldsymbol{h}) = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i - \boldsymbol{x}_j = \boldsymbol{h}, i, j = 1, 2, 3, \ldots, n\}$, which is unbiased (see [35, p. 69]).

The estimator $\widehat{\gamma}(\boldsymbol{h})$ in (14.6) cannot be used to compute $\boldsymbol{\lambda}$ (and therefore, it cannot be used to compute predictions $p(Y(\boldsymbol{x}_0)))$ since it may be negative definite, and it may thus yield non-sensical negative MSE's (see [35]). Common practice is to fit a parametric model to $\widehat{\gamma}(\boldsymbol{h})$ using least squares (which has the advantage of not depending on the distribution assumptions, as MLE's do). The form of the parametric model ensures that the resulting estimator $\tilde{\gamma}(\boldsymbol{h})$ is positive definite.

One useful parametric semi-variogram model is the exponential model:

$$\tilde{\gamma}(\boldsymbol{h}) = \begin{cases} c_0 + c_1(1 - exp(-||\boldsymbol{h}||/a)) & \text{if } \boldsymbol{h} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{h} = \mathbf{0} \end{cases}$$

where c_0 is called the *nugget* (recall the mining origin of Kriging!), $c_0 + c_1$ is called the *sill*, and a is called the *range*. This is an instance of an *isotropic* variogram model, which means it is assumed that the covariance between two points depends only on their separation or "lag" $||\boldsymbol{h}|| = ||\boldsymbol{x}_0 - \boldsymbol{x}_1||$ but not in the direction of $\boldsymbol{h}$. The nugget is supposed to capture discontinuities at the origin[4] caused by "micro" variation, which in practice occur due to measurement error.

Example. Kriging interpolation of a complicated 2-variable function. Suppose the (unknown) function

$$Y(\boldsymbol{x}) = x_1^8 + x_2^5 - x_1^2 x_2^2$$

represents the output of a computer code which is of interest to model as a function of two controllable factors.

The function is plotted in Figure 14.1. A space filling design (see Section 14.2) in the form of a grid $-.8 \leq x_1, x_2 \leq .8$ in increments of 0.4 was used

[4]That is, $\tilde{\gamma}(\boldsymbol{h}) \to c_0$ as $\boldsymbol{h} \to \mathbf{0}$.

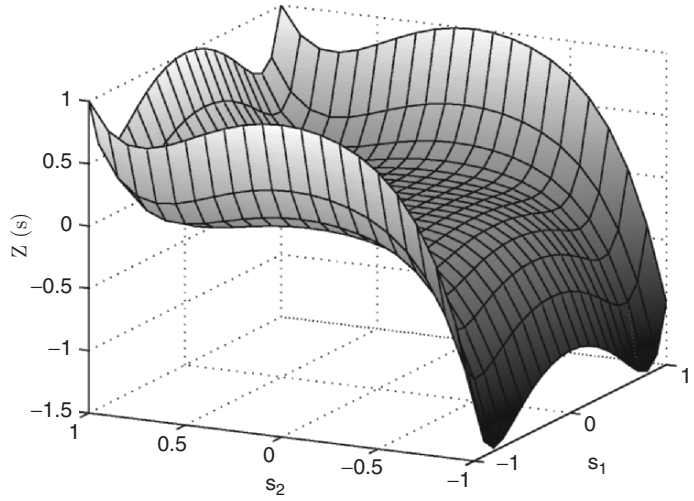


Figure 14.1. Contour plot of the true function $Y(\boldsymbol{x})$

to run experiments to get measurements of $Y(\boldsymbol{x})$. From these measurements, $\widehat{\gamma}(\boldsymbol{h})$ was estimated[5], see Figure 14.2. A least squares fit of the exponential model gives $c_0 = 0.02318, c_1 = 1.5255$, and $a = 0.95$. A contour of the predictions, obtained from (14.2) and from the fitted exponential semi-variogram is shown in Figure 14.3. Figure 14.5 displays the variance of the predictions, which, not surprisingly, increases inside the "pockets" left by the DOE points. Finally, Figure 14.4 gives the observed $Y(\boldsymbol{x})$ compared to the predictions obtained by deleting observation i when this is the point being predicted. The predictions are, in general, adequately following the shape of the true response surface. ■

14.1.1 Other Kriging Methods

A much more flexible model than (14.1) is obtained by assuming that the process mean μ is a function of the controllable factors, namely,

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \delta(\boldsymbol{x}).$$

[5]In this example, the Matlab software Easykrig v. 3.0 was used.

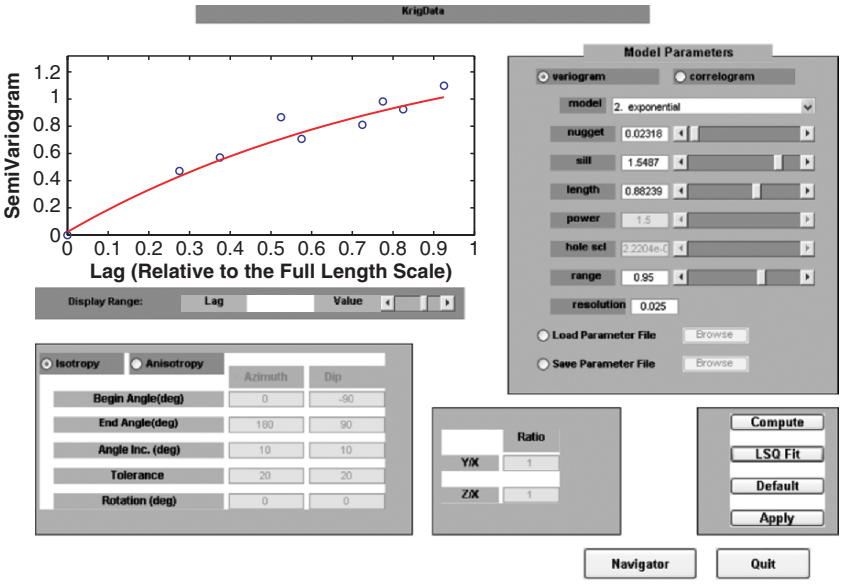


Figure 14.2. Fitted Semi-variogram $\widehat{\gamma}(\boldsymbol{h})$, exponential model

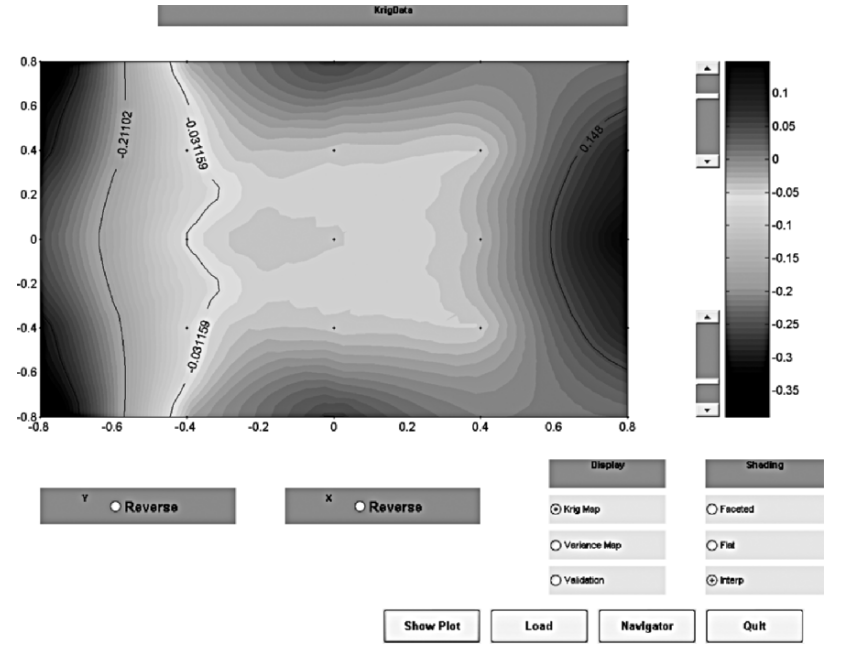


Figure 14.3. Contour plot of the predictions $p(Y(\boldsymbol{x}_0)$ for $-8 \leq x_1, x_2 \leq .8$

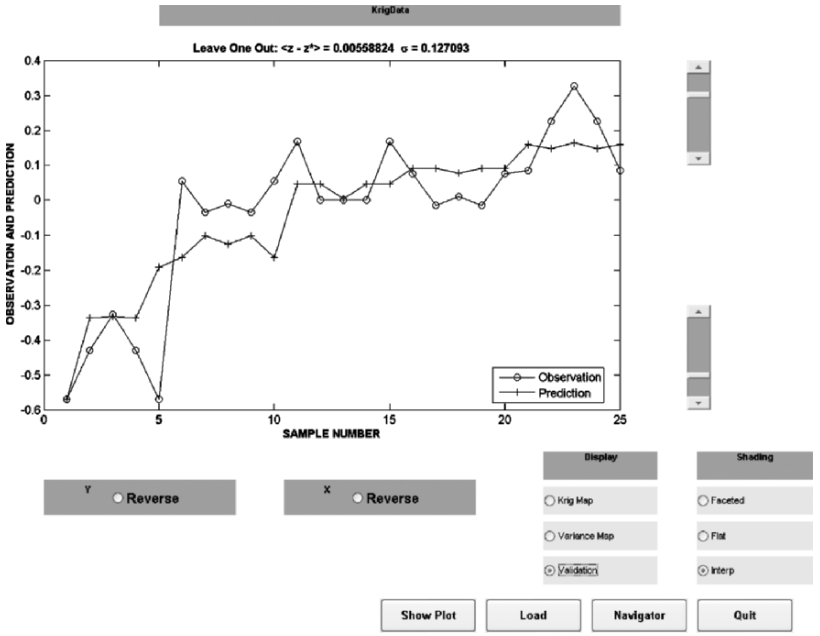


Figure 14.4. Validation of the predictions (observed vs. predicted) if predictions are made at deleted points

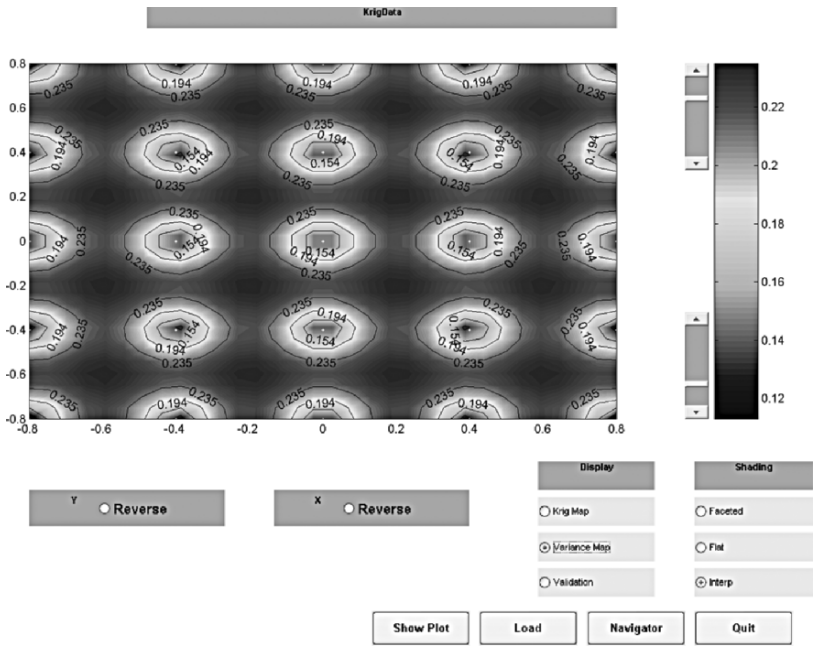


Figure 14.5. Variance plot of the predictions $p(Y(\boldsymbol{x}_0))$

Matheron (see [35]) studied such a model (called "Universal Kriging") with

$$\mu(\boldsymbol{x}) = \sum_{j=1}^{p+1} f_{j-1}(\boldsymbol{x})\beta_{j-1}$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ is an unknown vector of parameters, and $\boldsymbol{f}' = (f_0(\boldsymbol{x}), f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x}))$ is a vector of known functions. With this setting, the data $\boldsymbol{Y}$ can then be written as

$$\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\delta}$$

where $\boldsymbol{F}$ is a $n \times (p+1)$ matrix whose (i,j) element is $f_{j-1}(\boldsymbol{x}_i)$.

The form of the predictor is

$$\widehat{Y}(\boldsymbol{x}_0) = \sum_{i=1}^{n} \lambda_i Y(\boldsymbol{x}_i)$$

with $\boldsymbol{\lambda}'\boldsymbol{F} = \boldsymbol{f}'$. Note how if $p = 0$ and $f_0(\boldsymbol{x}) = 1$, the latter condition reduces to $\sum_{i=1}^{n} \lambda_i = 1$.

Optimal linear unbiased estimators for Universal Kriging and the associated problem of how to estimate the variogram for this model are treated by Cressie ([35, Chapter 3]). The MATLAB Kriging toolbox DACE [86] allows for Universal Kriging modeling.

A simpler alternative to Universal Kriging has been proposed by Van Beers and Kleijnen [153] who called their method *Detrended Kriging*. This can be used with only ordinary Kriging methods. In this approach, the mean is assumed to follow the model

$$\mu(\boldsymbol{x}) = S(\boldsymbol{x}) + \eta$$

where $S(\boldsymbol{x})$ is a known linear regression model (the "signal") of the form $S(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$, and η is white noise, i.e., $E[\eta] = 0$ and $\mathrm{Var}(\eta)$ is a constant. Under these assumptions, ordinary least squares can be used to estimate $S(\boldsymbol{x})$ from the data $\{\boldsymbol{x}_i, Y(\boldsymbol{x}_i)\}_{i=1,\ldots,n}$. The next step in this method is to apply Ordinary Kriging to the detrended data set

$$\{\boldsymbol{x}_i, Y(\boldsymbol{x}_i) - \widehat{S}(\boldsymbol{x}_i)\}_{i=1,\ldots,n}.$$

Finally, the prediction at point $\boldsymbol{x}_0$ is given by the *sum* of the OLS prediction, $\widehat{S}(\boldsymbol{x}_0) = \widehat{\boldsymbol{\beta}}'\boldsymbol{x}_0$, and the ordinary Kriging prediction, i.e.,

$$\widehat{Y}(\boldsymbol{x}_0) = \widehat{S}(\boldsymbol{x}_0) + p(Y(\boldsymbol{x}_0) - \widehat{S}(\boldsymbol{x}_0)).$$

Van Beer and Kleijnen report good performance of their proposal, although no direct comparisons with Universal Kriging were provided.

14.2 Space Filling Designs

In computer experiments, an experimental design $\boldsymbol{X}$ (an $n \times k$ matrix) specifies the n different computer runs, each at a particular set of k factor values. As mentioned earlier, for deterministic computer codes, the main problem is how to deal with the only source of uncertainty, namely, that associated with the unknown response function(s).

Computer experiments are frequently conducted for optimization purposes. For optimization, it is vital that a good representation of the unknown function be available over *the whole* region of exploration. This implies that "the region of experimentation" equals the region of interest (see Chapter 1), and, contrary to RSM, models are not local but *global*, fitted perhaps using Kriging techniques, as discussed in the previous section. Here we give an introduction to the main types of DOEs useful for deterministic computer experiments. For a more detailed discussion see [139].

In order to reduce the bias of the functional approximation, a design that places points evenly over the region of interest is desirable. We refer to such designs as *space filling designs*[6]. We note that a peculiarity of space filling design for computer experiments is that replication is unnecessary as the output of the computer code is deterministic.

Assuming a cuboidal region of interest, some obvious choices of space filling designs that are easy to implement are:

1 **Random designs.** These designs consist of n points generated at random within the region of interest.

2 **Random stratified designs.** the n points are obtained by partitioning the region of interest in n strata that covers the region. The one random point is selected from within each strata.

3 **Grid designs.** each factor is tried at m equidistant levels and all factorial combinations are tried, resulting in the $n = m^k$ points.

[6]One way of being more precise about what we mean by a space-filling design is to measure the discrepancy between a k-dimensional uniform distribution and the empirical distribution of the design. See [139] for details.

The disadvantages of each of these designs are clear. Random designs may not always cover the region of interest, even in the stratified form. Grid designs may require a considerable number of runs to cover a region adequately for a particular application.

Latin Hypercube sampling is a popular method to create DOEs in computer experiments, so we describe them next.

14.2.1 Latin Hypercube Designs (LHD)

These designs are an extension of stratified designs in which a modification is added as an effort to cover the region more evenly. LHDs are constructed with the following method:

1 Partition each of the ranges of the k factors into n intervals (usually, these are equally spaced, but if a subregion is of more interest, narrower intervals can be defined in it.) This will give a total of n^k "cells". Initially, mark all cells as "free".

2 From al the free cells, select one at random, say cell c. Pick a random point inside this cell (i.e., $x \in c$).

3 Mark all cells that agree with c on any of its coordinates as "used", i.e., cross out (mark as used) all cells in the same row, column, etc. as cell c.

4 If n cells have been selected, stop. Otherwise, go to step 2.

LHDs have nice *marginal* properties, that is, they spread points evenly across the whole range of each factor x_i. However, the distribution in n-dimensional space may not cover the region evenly enough.

Example. LHD design. Suppose we wish to construct a 10-point LHD for 2 factors, where $0 < x_1, x_2 < 1$. Figure 14.6 shows the resulting design. ■

LHDs have the obvious disadvantage, shared with all the simpler random designs, of not always being space filling in k-dimensional space. For example, for $k = 2$ all n observations could fall along the diagonal. Because of this, LHDs are better used in generating candidate designs from where a second space-filling criterion is used to make the final selection. One useful

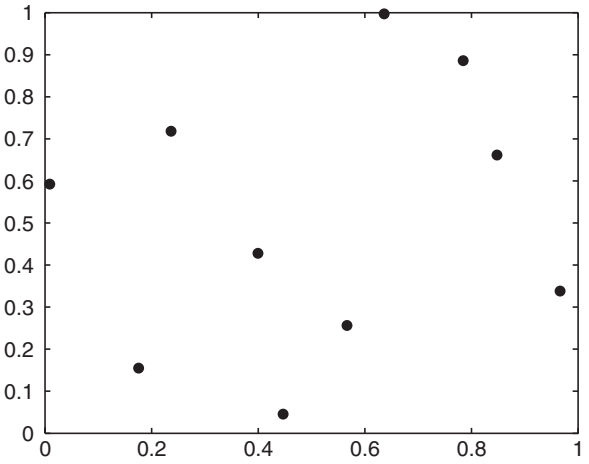


Figure 14.6. A Latin Hypercube Design for $k = 2, n = 10$

space-filling design is the *maximim* distance criterion. If the distance between two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is measured by the L_2 norm, i.e.,

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2|| = \sum_{i=1}^{k} (x_{1,i} - x_{2,i})^2$$

then a maximim design attempts to maximize the shortest distance between any two points of a design D:

$$\max \min_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in D} \; d(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

where the outer maximization is over all possible designs. This criteria can be combined with LHD by performing the maximization over a set of LHD designs:

$$\max_{D \in \text{LHD}(m,k,n)} \; \min_{\boldsymbol{x}_1, \boldsymbol{x}_2 \in D} \; d(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

where LHD(m, k, n) is a set of m randomly generated LHD designs in k dimensions and n points.

Matlab's Statistics Toolbox provides the function `lhsdesign` which can compute maximim LHDs. Its syntax is:

```
X=lhsdesign(n,k,'criterion','name','iterations',m)
```

where writing 'maximim' for 'name' causes the function to return the best maximim design among the m randomly generated LHD designs for the specified k and n.

Example. Maximim LHD. Figure 14.7 shows the maximim design returned by Matlab when the `lhsdesign` function was used for the case $n = 10$, $k = 2$, and $m = 500$. Figure 14.9 shows another maximim LHD obtained for $k = 3$, $n = 30$ and $m = 1000$. This design is given also in Table 14.1. ■

Matlab's `lhsdesign` function also provides the minimum correlation criterion (specify 'correlation' under 'criterion'), in which the LHD design with the least cross-correlations among the columns of the design is selected. If ρ_{ij} denotes the correlation between columns i and j, then this criterion is to choose the LHD design which

$$\min \sum_{i=1}^{k} \sum_{j>i} \rho_{ij}^2$$

that is, this criterion minimizes the sum of the squares of the elements in the upper triangular submatrix of the correlation matrix.

Example. Minimum Correlation LHD design. Figure 14.8 shows the LHD design with minimum sum of squared correlations for the case $k = 2$, $n = 10$, $m = 500$. The correlation between the two columns in the design is $\rho_{i2} = -0.0303$. In contrast, the designs in Figures 14.4 and 14.6 have

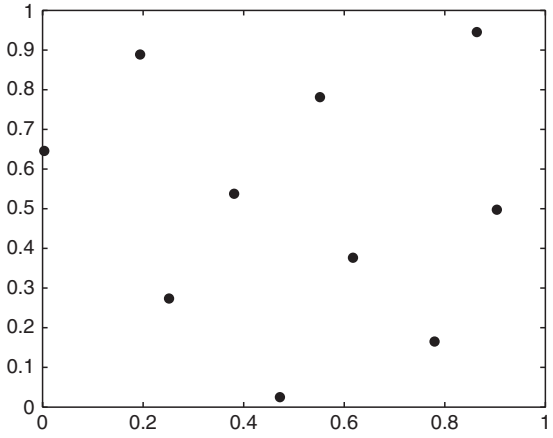


Figure 14.7. The best of 500 Latin Hypercube Designs selected according to the maximin criterion, $k = 2, n = 10$

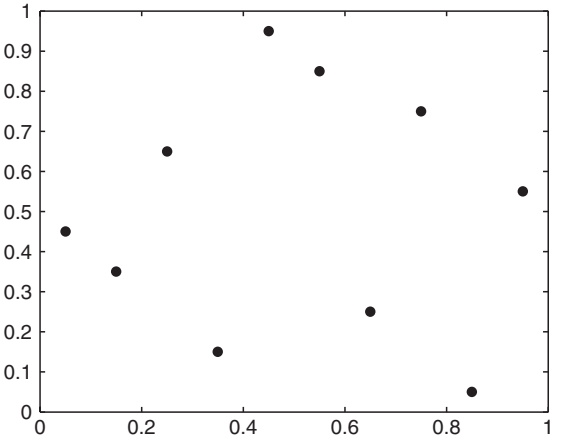


Figure 14.8. The best of 500 LHDs selected according to the correlation criterion, $k = 2$, $n = 10$

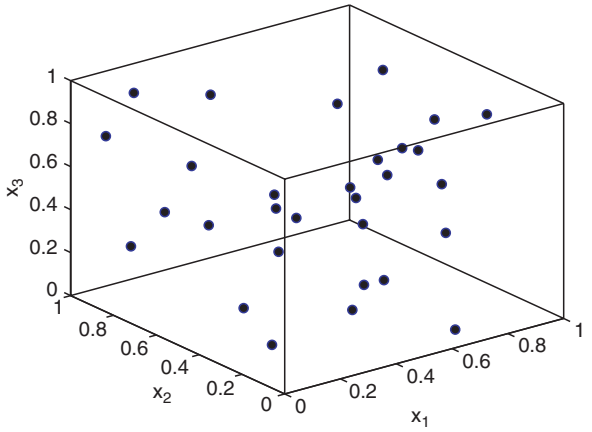


Figure 14.9. The best of 1000 LHDs selected according to the maximin criterion, $k = 3$, $n = 30$. The design points are tabulated in Table 14.1

$\rho_{12} = 0.2062$ and $\rho_{12} = -0.0715$ respectively. The maximin LH design in Figure 14.9 has $\rho_{12} = -0.0606, \rho_{13} = 0.0119$, and $\rho_{23} = 0.0307$.■

Another useful criterion (not available in Matlab) is to select the LHD design that minimizes the discrepancy between the CDF of a k-dimensional uniform distribution and the empirical distribution function of the LHDs. This is probably the best way to measure how "evenly" distributed the design points are in n-dimensional space. See [139] for more details.

Table 14.1. LHD design selected according to the maximin criterion, $k = 3, n = 10$ (best of 1,000 designs found). The points are plotted in Figure 14.9

x_1	x_2	x_3
0.83499	0.63006	0.92565
0.65383	0.10002	0.47383
0.80295	0.74044	0.34035
0.24747	0.66924	0.99873
0.99496	0.81773	0.29391
0.53273	0.44827	0.95713
0.16312	0.27186	0.04758
0.04598	0.89490	0.77358
0.34713	0.50045	0.57538
0.09490	0.31505	0.22225
0.70612	0.18602	0.64322
0.22498	0.99756	0.86466
0.38874	0.86233	0.25347
0.01655	0.06310	0.82888
0.56870	0.37545	0.41761
0.16855	0.93766	0.19895
0.88678	0.21115	0.89353
0.42405	0.23683	0.13323
0.56503	0.36589	0.14200
0.76731	0.56474	0.56101
0.26825	0.78407	0.60739
0.67125	0.90397	0.01153
0.62966	0.02369	0.06788
0.92003	0.49889	0.72595
0.46534	0.14290	0.30186
0.13233	0.73322	0.46322
0.96488	0.63356	0.50532
0.48463	0.57789	0.38316
0.33041	0.09804	0.75094
0.73916	0.41550	0.69297

14.3 Problems

1 Create a 50 point LHD maximim design for two variables x_i each with range $(-1,1)$ by using Matlab. Specify 500 iterations. Compute the correlation matrix of the design columns.

2 Suppose a computer program provides output of two variables that is described by the function $8x_1^4 + x_2^3$. "Sample" this function at each of the points in the design of Problem 1.

3 Perform ordinary kriging with the sampled values in Problem 2. In particular:

a) Using an exponential semi-variogram, give the fitted values of c_0, c_1, and a

b) Show the plot of the kriging approximation

c) Validate the fitted model using the "deleted" predictions.

(Hint: use the EasyKrig Matlab program).

4 Create a 50 point LHD minimum correlation design for two variables x_i each with range $(-1,1)$ by using Matlab. Specify 500 iterations. Compute the correlation matrix of the design columns. Compare with the design in Problem 1.

5 Sample the function in Problem 2 using the LHD min. correlation design of Problem 4.

6 Repeat Problem 3 using the LHD minimum correlation design and the function data of Problems 4–5. Which kriging approximation seems better?

7 Create a 100 point LHD maximim design for three variables x_i each with range $(-1,1)$ by using Matlab. Specify 500 iterations. Compute the correlation matrix of the design columns.

8 Suppose a computer program provides output of three variables that is described by the function $x_1^5 + x_2^5 + x_3^5 + x_1x_2 + x_2x_3 + x_1x_3$. "Sample" this function at each of the points in the design of Problem 7.

9 Perform ordinary kriging with the design and sampled values in Problems 7 and 8. In particular:

a) Using an exponential semi-variogram, give the fitted values of c_0, c_1, and a

b) Show the plot of the kriging approximation

c) Validate the fitted model using the "deleted" predictions.

(Hint: use the EasyKrig program).

10 Create a 100 point LHD minimum correlation design for three variables x_i each with range (−1,1) by using Matlab. Specify 500 iterations. Compute the correlation matrix of the design columns. Compare with the design in Problem 7.

11 Sample the function in Problem 8 using the LHD min. correlation design of Problem 10.

12 Repeat Problem 9 using the LHD minimum correlation design and the function data of Problems 10–11. Which kriging approximation seems better?

13 Derive (14.4) from (14.3). Work out the binomial term and use the fact that $\sum_{i=1}^{n} \lambda_i = 1$.

14 Derive (14.5). Hint: use the result:

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{F}\boldsymbol{E}^{-1}\boldsymbol{F}' & -\boldsymbol{F}\boldsymbol{E}^{-1} \\ -\boldsymbol{E}^{-1}\boldsymbol{F}' & \boldsymbol{E}^{-1} \end{pmatrix}.$$

PART VII

APPENDICES

Appendix A
Basics of Linear Regression

This appendix provides an overview of linear least squares estimates, model diagnostics, and some tests of hypothesis relevant in response surface methods.

A.1 Ordinary Least Squares

In most experimental designs, the following information is collected from an experiment:

$$\begin{array}{ccccc} y_1 & x_{11} & x_{12} & \cdots & x_{ik} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ y_3 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nk} \end{array}$$

where n is the number of experimental runs and k the number of controllable factors. Thus, for each observed response, we also measure and record the process settings under which the observation was obtained, i.e., the levels of the controllable factors. Here it is assumed that all controllable factors are quantitative; otherwise, some of the regressors need to be indicator variables[1].

Suppose we postulate a first order model as being true for all the n observations:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, 2, \ldots, n.$$

The **least squares criterion** finds estimates $\widehat{\beta}_0, \widehat{\beta}_1, \ldots \widehat{\beta}_k$ such that we minimize the sum of squared residuals:

$$R(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k) = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right)^2 .$$

[1]To represent a categorical factor with k categories, $k-1$ zero-one indicator variables are necessary.

The function is convex, so its minimum is obtained from $\frac{\partial R}{\partial \beta_j} = 0, j = 0, .., k$. This gives a system of $p = k + 1$ equations in $p = k + 1$ unknowns, where p denotes the number of parameters, which equals in this particular case the number of controllable factors plus one. Unless we have chosen our controllable factor levels badly (something which will be made clear shortly), this system of equations has a unique solution.

The matrix representation of the first order model and data is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

which includes the experimental design and is a $n \times p$ matrix in general,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

which is $p \times 1$ in general, and

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The sum of squared residuals is:

$$R(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The normal equations are obtained from $\frac{\partial R(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} = \mathbf{0}$ and yield

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}.$$

The least squares estimate of $\boldsymbol{\beta}$ is

$$\boldsymbol{b} = \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{A.1}$$

where

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ik} \\ & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ & & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots \\ \text{symmetric} & & & & \sum x_{ik}^2 \end{bmatrix}$$

is a $p \times p$ matrix of sum of crossproducts. The matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is also symmetric. This inverse exists if $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}') = p$, i.e., if $\boldsymbol{X}$ is of full rank.

The residuals are obtained from the predicted observations:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \equiv \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is a $n \times n$ symmetric matrix called the "hat" matrix in the regression literature, a term apparently coined by by John Tukey (so that we can remember that "y hat equals hat y"). The matrix $\boldsymbol{H}$ is a projection matrix (see Appendix C).

The residuals of the fitted model are $\widehat{\varepsilon}_i \equiv e_i = y_i - \widehat{y}_i$, or

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

where the matrix $\boldsymbol{I} - \boldsymbol{H}$ is also a projection matrix.

Figure A.1 shows a geometric interpretation of the least squares residual. The vector $\boldsymbol{Y}$ is in n-dimensional space, and the space spanned by the columns of $\boldsymbol{X}$ is a subspace of it, where any vector of the form $\boldsymbol{X}\boldsymbol{\beta}$ lies. The least squares criterion proposes to choose $\widehat{\boldsymbol{\beta}}$ such that the euclidean distance from $\boldsymbol{Y}$ to $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is as small as possible. This in turn results in the length of vector $\boldsymbol{e}$ to be the shortest possible one.

A closer look at the triangle formed by the three vectors $\boldsymbol{Y}$, $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ indicates that that $\boldsymbol{e}$ is orthogonal to the space spanned by $\text{col}(\boldsymbol{X})$. This means that

$$\boldsymbol{X}'\boldsymbol{e} = 0 = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

from which we obtain the normal equations:

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$$

and solving we get the **ordinary least squares (OLS)** estimate of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.$$

Furthermore,

$$\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \boldsymbol{e}$$

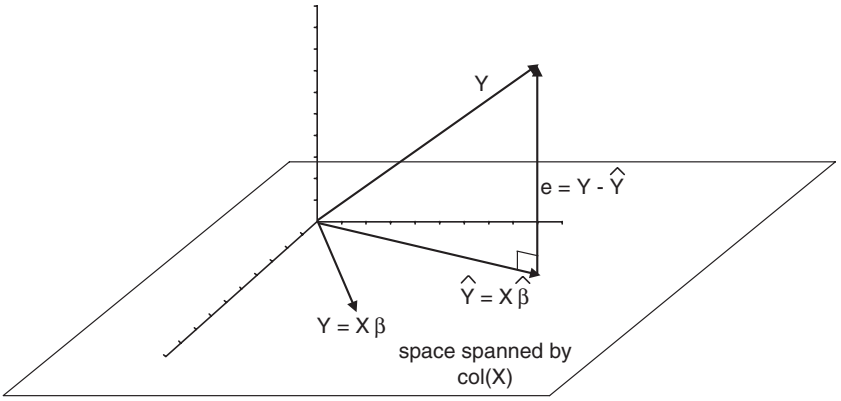


Figure A.1. Geometric interpretation of least squares

thus we have partitioned the vector of observations into a vector on the space spanned by col($\boldsymbol{X}$) and a vector on the space orthogonal to col($\boldsymbol{X}$). The proof of the orthogonality of $\boldsymbol{e}$ and $\widehat{\boldsymbol{Y}}$ uses the symmetry and idempotency properties (see Appendix C) of the hat matrix. We have that:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}, \quad \text{and} \quad \boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

Therefore,

$$\begin{aligned} \widehat{\boldsymbol{Y}}'\boldsymbol{e} &= \boldsymbol{Y}'\boldsymbol{H}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{H}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\ &= \boldsymbol{Y}'(\boldsymbol{H} - \boldsymbol{H}^2)\boldsymbol{Y} = 0. \end{aligned}$$

We thus can see that the $\boldsymbol{H}$ matrix projects a vector onto the space spanned by col($\boldsymbol{X}$) while the matrix $(\boldsymbol{I} - \boldsymbol{H})$ projects a vector onto the space orthogonal to that spanned by col($\boldsymbol{X}$). The orthogonality of $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{e}$ is a consequence of certain assumptions made with respect to the errors ε_i in the model. Therefore, this orthogonality is exploited in regression diagnostics, where the objective is to check if the assumptions behind least squares estimation hold.

A.1.1 Properties of OLS Estimates

How good are the OLS estimates? Their "goodness" depends on certain assumptions being true. In addition to assuming the general model form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, there are three assumptions made in basic linear regression with respect to the errors:

1 $E[\boldsymbol{\varepsilon}] = \boldsymbol{0}$

2 Var$[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}$

3 $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$

If assumptions 1 and 2 hold, OLS are "BLUE" (Best Linear Unbiased Estimators). This means that $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$ have, individually, the smallest variances of all linear unbiased estimators of these parameters, a result known as the Gauss-Markov theorem (an estimator T is linear if it is

of the form $T = l_1 y_1 + l_2 y_2 + \cdots + l_n y_n$ where $l_1, l_2, \ldots, l_n$ are some constants. An estimator $\widehat{\theta}$ is unbiased for θ if $E[\widehat{\theta}] = \theta$).

If assumption 3 is also true, then errors are independent, not only uncorrelated, and the OLS are UMVUE (Uniformly Minimum Variance Unbiased Estimators), that is, they have the same properties than in the BLUE case but for all classes of estimators (linear and nonlinear), evidently a stronger result. Thanks to normality, we can make t and F tests of significance and construct confidence intervals about the parameters. In addition, if all 3 conditions hold, we can show that OLS estimators are MLE (maximum likelihood estimators), which have nice asymptotic properties. This is a statement easy to prove:

LEMMA A.1 *If conditions 1-3 hold, the OLS (A.1) are MLE's.*

Proof. The joint probability function of the errors is

$$p(\boldsymbol{\varepsilon}) \propto \frac{1}{\sigma^n}\, e^{-\frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{2\sigma^2}}.$$

Since $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the likelihood function is

$$L(\boldsymbol{\beta}|\boldsymbol{Y}) \propto \frac{1}{\sigma^n}\, e^{-\frac{\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\right)}{2\sigma^2}} = \frac{1}{\sigma^n}\, e^{-\frac{R(\beta)}{2\sigma^2}}$$

which is maximized with respect to $\boldsymbol{\beta}$ when the sum of squares function $R(\boldsymbol{\beta})$ is minimized with respect to $\boldsymbol{\beta}$ ■.

We can also show the following.

LEMMA A.2 *The OLS's (A.1) are unbiased.*

Proof.

$$\begin{aligned} E[\boldsymbol{b}] = E[\widehat{\boldsymbol{\beta}}] &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{Y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \boldsymbol{\beta} \ \blacksquare. \end{aligned}$$

The variance-covariance matrix of the parameter estimates is:

$$\begin{aligned} \text{Var}[\boldsymbol{b}] &= \text{Var}(\widehat{\boldsymbol{\beta}}) = \text{Var}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\text{Var}\ \boldsymbol{Y}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}. \end{aligned}$$

Therefore, apart from a constant, the diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ contain the variances of each $\widehat{\beta}_j$ and the off-diagonal elements contain the covariances of the (β_i, β_j) pairs. Notice how designing an $\boldsymbol{X}$ such that $(\boldsymbol{X}'\boldsymbol{X})$ (and $(\boldsymbol{X}'\boldsymbol{X})^{-1}$) are diagonal makes sense; these are the so-called orthogonal designs.

Since the errors ε_i are normal, the observations are normal. The OLS are also normal, as they are linear combinations of the observations. Putting together the previous results we have that

$$\boldsymbol{b} = \widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

A.1.2 Estimation of σ^2

If we wish to test hypothesis or compute confidence intervals on the β's, σ^2 needs to be estimated. If conditions 1-3 hold, then the UMVUE of σ^2 is

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p}(\boldsymbol{Y}-\boldsymbol{Xb})'(\boldsymbol{Y}-\boldsymbol{Xb})$$
$$= \frac{\boldsymbol{e}'\boldsymbol{e}}{n-p} = \frac{R(\boldsymbol{b})}{n-p}$$

where $n-p$ are the degrees of freedom for error. This is not the MLE estimator, which has a denominator of n instead and is biased.

Note that if biased estimators are allowed, then there might be estimators of $\boldsymbol{\beta}$ and σ^2 that may have lower MSE than the OLS estimators. Furthermore, despite the BLUE property of the OLS estimates when conditions 1-2 hold, OLS estimates are not robust to non-normal errors. Evidently, an abnormally large residual will have a large impact in the sum of squares, and it may "lever" the fit in its favor, becoming a very influential observation. Thus, lack of normality and abnormally large errors are a concern.

A.1.3 Estimation and Prediction in Regression

Suppose we fit a first order model. Let $\boldsymbol{x}_0' = (1, x_{01}, x_{02}, \ldots, x_{0k})$ be a point of interest where we wish to investigate the response. Then, $\widehat{y} = \widehat{Y}(\boldsymbol{x}) = \boldsymbol{x}_0'\widehat{\boldsymbol{\beta}} = \boldsymbol{x}_0'\boldsymbol{b}$ is a *point* estimator of both of the following:

1 the mean of Y at $\boldsymbol{x} = \boldsymbol{x}_0$, i.e., $E[y|\boldsymbol{x}_0]$;

2 the next observation of y at $\boldsymbol{x}_0$, i.e., $y(\boldsymbol{x}_0)$,

Note how these two quantities are fundamentally different. In classical ("frequentist") statistics, $E[y|\boldsymbol{x}_0]$ is an unknown constant. In contrast, the next observation at $\boldsymbol{x}_0$ is a random variable which is distributed as $N(\boldsymbol{\beta}'\boldsymbol{x}_0, \sigma^2)$.

Confidence intervals for each of these quantities are different. For the 1st case, a confidence interval for $E[Y|\boldsymbol{x}_0]$ is given by

$$\widehat{y}(\boldsymbol{x}_0) \pm t_{\alpha/2,n-p}\,\widehat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

In the second case, a **prediction interval** for $Y(\boldsymbol{x}_0)$ is given by

$$\widehat{y}(\boldsymbol{x}_0) \pm t_{\alpha/2,n-p}\,\widehat{\sigma}\sqrt{1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}. \tag{A.2}$$

Here, the quantity $\sqrt{\text{Var}(\widehat{y}(\boldsymbol{x}_0))} = \sqrt{\text{Var}(\boldsymbol{x}_0'\boldsymbol{b})} = \sqrt{\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$ is used in case 1, and $\sqrt{\text{Var}(y(\boldsymbol{x}_0) - \widehat{y}(\boldsymbol{x}_0))} = \sqrt{\sigma^2 + \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$ (standard error of the prediction), is used in case 2. Note how the latter contains the variance of the future observation (σ^2) *plus* the variance associated to our point estimate. Thus, the prediction interval will always be wider than the confidence interval for the mean response. Finally, also note that the interval estimates depend on both the experimental design and on the point on the controllable factor space at which the estimate is desired.

A.1.4 Analysis of Residuals and Influence Diagnostics

We use the residuals $e_i = y_i - \widehat{y}_i$ to determine if the assumptions made on the errors ε_i hold true. The expected value of a residual is zero, since $E[\boldsymbol{e}] = E[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}] = \boldsymbol{0}$ (here one has to remind the classical statistics concept that a statistic such as $\boldsymbol{e}$ is a quantity based on random data, and therefore is a random variable itself *prior* to observing the data). The variance of the residuals is

$$\text{Var}(\boldsymbol{e}) = \text{Var}((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{H})\text{Var}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{H})' = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

because $(\boldsymbol{I} - \boldsymbol{H})$ is idempotent. Thus, the variance of the ith residual is not a constant:

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

where h_{ii} is the ith diagonal element of matrix $\boldsymbol{H}$. Compare this with the assumption of constant variance on $\boldsymbol{\varepsilon}$. The variance of the residuals depends on the experimental design. Similarly,

$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij},$$

compared to no correlation between the ε_i's. The relation between residuals and errors is:

$$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$$

thus we can see that each residual is a function of all errors ε_j.

All of this implies that the ordinary residuals have different variances and are correlated, even when the assumptions on the errors (constant variance, no correlation) are true. Because of this, is it customary to transform the residuals to make their variance constant defining:

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n$$

which has a constant variance of one. This are sometimes called "standardized" or "studentized" residuals, depending on the author. Their distribution is close to a Student t with $n - p$ degrees of freedom, but is not exactly t distributed. A problem with it is that if observation i is an outlier, s will overestimate σ. A solution is to use instead the "R-student" residuals (called studentized residuals or studentized deleted residuals, depending on different authors):

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n$$

where

$$s_{(i)} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1 - h_{ii})}{n - p - 1}}, \quad i = 1, \ldots, n$$

is the standard deviation estimated from the data set after deleting the ith data point. t_i will differ from r_i if observation i is influential. The R-student statistics are distributed as a Student t with $n - p - 1$ degrees of freedom[2]. For n larger than p, a good rule of thumb to determine if an observation is an outlier or not is that $|t_i| > 2$.

[2]Here we point out that the t_i's (and the r_i's) may still be correlated.

Not all outliers are influential. Influence depends also on *leverage*. To assess the leverage of point i on $\widehat{y}$, compute $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$, the ith element on the diagonal of the $\boldsymbol{H}$ matrix. The h_{ii} values depend on the Euclidean distance from the centroid of all the design points to point $\boldsymbol{x}_i$ and on how dense the points are in the region where point i lies. The smallest convex set containing all the design points is called the independent variable hull (IVH, see [32]). For all $\boldsymbol{x}$ in the IVH ($\boldsymbol{x}$ need not be a design point),

$$\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} \leq \max_i h_{ii}.$$

This follows because the left hand of the above equation are ellipsoids, and the ellipsoid passing through the design point associated to max h_{ii} must contain all the IVH. This also implies, in view of (A.2), that the point with the largest variance of a prediction must lie on the boundary of the IVH. This point needs not be the point farthest from the centroid; as said before, the density of points also matters. In general, the design point with largest h_{ii} value will be on the boundary of the IVH in a region where the density of design points is relatively low. Predicting at a non-design point such that $\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} > \max h_{ii}$ will imply an extrapolation outside the region where the DOE was conducted on the controllable factor space.

The quantities h_{ii} are usually referred to as the **"leverage"** of a point. A design point with large leverage will be located generally far from the area where most other design points are located. Such point will therefore have the *potential* to "lever" the regression equation, pulling it towards its y-value, if included in the data set (note that other definitions of influence exist). An observation is **influential** if the parameter estimates change depending on whether or not the observation is included in the data set. In general, observations associated with points with large h_{ii} (high leverage) and a large R-student residual (t_i) will be influential. A rule of thumb found in regression books to determine if leverage is large is to compare all the h_{ii} values against its average value, $\overline{h}$. It can be shown that $\overline{h} = p/n$. The rule of thumb calls point i a point of high leverage if $h_{ii} > 2\overline{h} = 2p/n$.

A useful statistic to determine influence is Cook's "distance", defined by:

$$D_i = \frac{(\boldsymbol{b} - \boldsymbol{b}_{(i)})'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{b} - \boldsymbol{b}_{(i)})}{p\ s^2}, \quad i = 1, 2, \ldots, n$$

where $\boldsymbol{b}_{(i)}$ is a vector of OLS estimates obtained after deleting observation i. D_i is a standardized measure of the distance between the $\boldsymbol{b}$ and $\boldsymbol{b}_{(i)}$ vectors, and gives a measure of the influence of observation i. An easier and more revealing computing formula is

$$D_i = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

which indicates that D_i tends to be large when r_i is large and when the distance to the centroid of point i, measured by $h_{ii}/(1 - h_{ii})$, is also large. The D_i statistics are distributed as an F with p and $n - p$ degrees of freedom. A rule of thumb useful when $n \gg p$ is that if point i has $D_i > 1$, then it should be considered an influential point. Influential points need to be investigated, and a non-statistical decision needs to be made (perhaps together with the engineers familiar with the process) as to whether one should include it in the analysis or not.

A.1.5 Diagnostic Plots

Besides computing residuals and determining influential observations, it is also customary in practice to plot some simple graphs that can be used to check the OLS assumptions. The most important such diagnostic plots are:

- Normal probability plot (NPP) of the residuals. Here we plot the e_i's to assess normality. If $n - p$ is large, we could plot the r_i's or the t_i's. Normality is needed to make statistical inferences like t and F tests of significance and to build interval estimates. It is not necessary for computing the OLS. However, the OLS are non-robust with respect to outliers, i.e., one largely influential observation will "break them down". Assessing normality from little data (as in many small DOE's) is very difficult, because the NPP itself has sampling error. It is possible to plot normally distributed data by simulation and get very non-normally-looking NPP's. If data is clearly non-normal, robust estimation techniques should be used instead of OLS. Bootstrapping techniques can also be used for non-normal inference [53].
- Residual vs. fitted values. This plot is used to check that the variance of the errors is constant. We plot t_i vs. $\widehat{y}_i$, expecting to see no relation since $\boldsymbol{e}$ (and $\boldsymbol{t}$) is orthogonal to $\widehat{\boldsymbol{Y}}$. Data transformations are usually employed to deal with non-homogeneous variance.
- e_i (or t_i) vs. i. This is used to determine if errors are correlated. If they are, a time series technique can be applied to model them, capturing the residual correlation. In some cases, only pairs of certain residuals are correlated, and this may result in influential observations, since deleting one point i could make some other residual e_j to increase or decrease.
- Residuals vs. factors. Here one plots the e_i (or t_i) vs. x_j, the values of factor j, to see if there is some higher order term (e.g., x^2, say) that should have been included in the model. This can also be done with respect to variables not in the model.

A.2 Testing Hypothesis on the OLS

A.2.1 Significance of Regression

The most basic test in regression, and perhaps the first thing to test, is

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{at least one } \beta_j \neq 0 \end{aligned}$$

To do this test, we perform an Analysis of Variance (ANOVA), which consists in partitioning the total sum of squares as follows:

$$\text{SS}_{total} = \text{SS}_{regression} + \text{SS}_{error}$$

where the sum of squared "errors" (in reality, the residuals), is

$$\text{SS}_{error} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

and since $\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}$,

$$\text{SS}_{error} = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

This formula measures deviations from the *origin*. It is common practice to measure deviations from the average response, namely, $y_i - \overline{Y}$ and assume that $\beta_0 \neq 0$, i.e., assume there is a non-zero intercept. The *corrected sum of squares* are then

$$\begin{aligned} \text{SS}_{total} &= (\boldsymbol{Y} - \overline{Y}\mathbf{1})'(\boldsymbol{Y} - \overline{Y}\mathbf{1}) = \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2 \\ \text{SS}_{regression} &= (\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1})'(\widehat{\boldsymbol{Y}} - \overline{Y}\mathbf{1})' = \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - n\overline{Y}^2 \\ \text{SS}_{error} &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \end{aligned}$$

The test statistic is

$$F_0 = \frac{\text{SS}_{reg}/(p-1)}{\text{SS}_{error}/(n-p)}$$

which follows a $F_{p-1,n-p}$ distribution under H_0. If H_0 is rejected, the variation modelled is significantly greater than the unexplained variation.

An important statistic related to this test is

$$R^2 = \frac{\text{SS}_{regression}}{\text{SS}_{total}}$$

which measures the proportion of total variation in the data explained by the model. Clearly, $0 \leq R^2 \leq 1$. This is a useful statistic to measure goodness of fit. However, R^2 can be artificially increased by either adding regressors (increasing p) or spreading the levels further apart. The effect of each of this is to reduce the elements in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, making $\text{SS}_{error} \to 0$ so $\text{SS}_{regression} \to \text{SS}_{total}$ and thus $R^2 \to 1$. Thus, comparing models with different number of parameters using the R^2 statistic is not recommended, as it will favor larger models. Instead, to compare models with different number of parameters it is better to use the adjusted R^2 statistic:

$$R^2_{adjusted} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right)$$

which does not necessarily increases as p increases since it takes into account the degrees of freedom lost. Using $R^2_{adjusted}$ to compare different models usually results in same conclusions as when comparing them based on the MS_{error}, which also considers the degrees of freedom.

Another useful statistics for model fit comparisons is the Predicted REsiduals Sum of Squares (**PRESS**) defined as

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \widehat{y}_{[i]})^2 = \sum_{i=1}^{n} e^2_{[i]}$$

where $\widehat{y}_{[i]}$ is the prediction given by a model fitted to all data *except* the ith observation y_i obtained at $\boldsymbol{x}_i$. Thus, this statistic is based on the idea of "leaving-one-out" (also called "jacknifing" technique). It is not necessary to refit the model to get $\widehat{y}_{[i]}$, since it turns out that

$$e_{[i]} = \frac{e_i}{1 - h_{ii}}$$

so PRESS is very easy to compute. A large value of PRESS indicates difficulties of the model for prediction (of the data observed, in this case). If PRESS is much larger than the usual sum of squared residuals, there may be outliers.

A.2.2 Tests on Individual and Subgroups of Parameters

Rejecting the null hypothesis in the test of significance of regression implies there are some non-negligible parameters, but it does not tell us which ones they are. A procedure for testing for the significance of any group of parameters is now described.

For example, suppose we want to see which of the two following models fits better:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{A.3}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{A.4}$$

We can set the test:

$$H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0 \quad \text{vs.} \quad \text{at least one } \beta_{ij} \neq 0.$$

For performing these tests, we use the **extra sum of squares principle** (ESS). We should:

1 fit the most complicated model first (this is the full or complete model);

2 fit the simpler (restricted or reduced) model

The ESS principle indicates that

$$\text{SS}_{regression,full} \geq \text{SS}_{regression,reduced}$$

where the difference $\text{SS}_{reg,full} - \text{SS}_{reg,reduced}$ is due to the additional (extra) parameters not in the reduced model. Intuitively, if this difference is large, then we should reject the null hypothesis (i.e., we conclude that the extra parameters are significant).

Let

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

where $\boldsymbol{\beta}_1$ has $p - r$ rows and $\boldsymbol{\beta}_2$ has r rows ($\boldsymbol{\beta}_1$ always contains β_0). The hypotheses are

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_2 \neq 0.$$

For the complete model, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and

$$\text{SS}_{regression,full} \equiv \text{SS}_{reg}(\boldsymbol{\beta}) = \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} - n\overline{Y}^2$$

which has $p - 1$ degrees of freedom (here, we use the notation $\text{SS}_{reg}(\boldsymbol{\beta})$ to emphasize the dependence on the full model), and

$$\text{SS}_{error}(\boldsymbol{\beta}) = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

which has $n - p$ degrees of freedom.

The reduced model is

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{X}_1$ is formed from the columns of $\boldsymbol{X}$ corresponding to the parameters in $\boldsymbol{\beta}_1$. The OLS of $\boldsymbol{\beta}_1$ is obtained in the usual form: $\widehat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}$ and

$$\text{SS}_{reg(\boldsymbol{\beta}_1)} = \boldsymbol{Y}'\boldsymbol{X}_1\widehat{\boldsymbol{\beta}_1} - n\overline{Y}^2$$

which has $p - r - 1$ degrees of freedom. The regression sum of squares due to $\boldsymbol{\beta}_2$ *given that* $\boldsymbol{\beta}_1$ is already in the model is:

$$\text{SS}_{reg}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \text{SS}_{reg}(\boldsymbol{\beta}) - \text{SS}_{reg}(\boldsymbol{\beta}_1)$$

which has $(p-1) - (p-r-1) = r$ degrees of freedom. This is the ESS, which measures the increase in SS due to $\boldsymbol{\beta}_2$. If H_0 is true then

$$F_0 = \frac{\text{SS}_{reg}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{\text{SS}_{error}(\boldsymbol{\beta})/(n-p)}$$

will be small. This statistic is distributed as a $F_{r,n-p}$ if H_0 is true.

If $r = 1$, we test for the individual significance of a single parameter. The corresponding tests are called "partial F tests". Since $F_{1,n-p} \equiv (t_{n-p})^2$, the tests are equivalent to t-tests on each parameter:

$$t_0 = \frac{\widehat{\beta}_j}{\widehat{\text{std.}}(\widehat{\beta}_j)}$$

where the standard error of each estimate is estimated by the square root of the diagonal elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix times s. It is important to note that these tests are for the significance of parameter β_j *given* that all other parameters are present in the model.

In RSM, we can use the ESS to test for curvature[3]. For example, suppose we wish to compare models (A.3) and (A.4). The ESS is given by

$$\text{SS}_{reg}(\beta_{12}, \beta_{11}, \beta_{22}|\beta_1, \beta_2, \beta_0) = \text{SS}_{reg}(\beta_1, \beta_2, \beta_{12}, \beta_{11}, \beta_{22}|\beta_0) - \text{SS}_{reg}(\beta_1, \beta_2|\beta_0)$$

from which we can test H_0.

The test for significance and the tests for subsets of parameters are particular cases of the general linear hypothesis (GLH):

$$H_0 : \boldsymbol{C}\boldsymbol{\beta}_p = \boldsymbol{w} \quad \text{vs.} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta}_p \neq \boldsymbol{w}$$

where $\boldsymbol{C}$ is a $r \times p$ matrix of rank r and $\boldsymbol{w}$ is a $r \times 1$ vector of constants. For example, if $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{w} = \boldsymbol{0}$, we get the significance of regression test. The test statistics in the GLH is

$$F_0 = \frac{(\boldsymbol{C}\widehat{\boldsymbol{\beta}_p} - \boldsymbol{w})'(\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}')^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{w})}{\text{SS}_{error}/(n-p)}$$

which follows a $F_{r,n-p}$ distribution under H_0.

A.2.3 Test for Lack of Fit

A fitted model may exhibit "lack of fit" (LOF) even if it is significant according to the significance of regression test and has a high R^2 statistic. The reasons for lack of fit are mainly two:

[3]This is *not* the single degree of freedom test of curvature shown later on.

1 Factors were omitted but affect the response;

2 Higher order terms of factors already in the model were omitted but affect the response

The null hypothesis we consider is:

$$H_0 : \text{there is } \textit{no} \text{ lack of fit.}$$

To test for LOF, an experimental design must satisfy the following requirements:

- the number of distinct experimental points, m, should be such that $m > p$, i.e., we should have a "non-saturated design";
- the design must have at least 2 replicates at one or more experimental point.

Let n_i be the number of replicates at point i. The second requirement is then $n_i > 2$ for some i. We have that

$$n = \sum_{i=1}^{m} n_i.$$

In this case, the ANOVA will proceed to partition the SS_{error}:

$$\text{SS}_{total} = \text{SS}_{reg} + \text{SS}_{error} = \text{SS}_{reg} + \text{SS}_{Pure\ Error} + \text{SS}_{LOF}.$$

Thus, the error sum of squares is partitioned as follows:

$$\begin{array}{ccccc} \text{SS}_{error} & = & \text{SS}_{PE} & + & \text{SS}_{LOF} \\ \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\widehat{y}_i)^2 & = & \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 & + & \sum_{i=1}^{m} n_i(\overline{y}_i-\widehat{y}_i)^2 \\ n-p & = & n-m & + & m-p \end{array}$$

where $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and the last row above shows the degrees of freedom. The SS_{error} measures variation in the residuals; the pure error sum of squares measures variation of the data around their averages. We note that this term does not involve any parameter estimate, thus it is *model independent*. As for the LOF term, the idea is that if there is no LOF, the predictions should be close to the average responses at each point, and therefore SS_{LOF} will be small. This provides a test on H_0, with test statistic:

$$F_0 = \frac{\text{SS}_{LOF}/(m-p)}{\text{SS}_{PE}/(n-m)}$$

which is distributed as a $F_{m-p,n-m}$ if H_0 is true. Notice that if a model fails the LOF test this implies "something is missing", but the test does not tell us what that is.

A.2.4 Single Degree of Freedom Test for Curvature

Depending on the DOE and the model to fit, the LOF sum of squares can be further partitioned. A practical case is when we use a 2^k factorial with center points to fit a 1st order model. Then

$$m - p = 2^k + 1 - (k+1).$$

From the binomial theorem, we have that $2^k = 1 + \sum_{i=1}^{k} \left(\begin{array}{c} k \\ i \end{array} \right)$, thus the degrees of freedom for LOF are

$$m - p = 2^k + 1 - (k+1) = 1 + \sum_{i=1}^{k} \left(\begin{array}{c} k \\ i \end{array} \right) + 1 - k - 1$$

or

$$m - p = 1 + \sum_{i=2}^{k} \left(\begin{array}{c} k \\ i \end{array} \right).$$

The "1" is a degree of freedom we can use to test for LOF due to "pure quadratic" terms that should be in the model (i.e., a test of curvature); the rest of the terms are the $m - p - 1$ degrees of freedom for all the possible interactions (of any order ≥ 2). Thus, we can split the LOF sum of squares as LOF due to each of these components.

Example. Test for curvature. Suppose $k = 2$ and we use a 2^2 factorial to fit a first order model. Thus, $m = 2^2 + 1 = 5$, $p = 3$, and $m - p = 2$ which can be split in one degree of freedom for curvature and one degree of freedom for the $x_1 x_2$ (or the "AB") interaction. ■

The partition of sums of squares and degrees of freedom in general is then

$$\begin{array}{lllllllllll} \text{SS}_{error} & = & \text{SS}_{PE} & + & \text{SS}_{LOF} & = & \text{SS}_{PE} & + & \text{SS}_{int.} & + & \text{SS}_{Curv.} \\ (n-p) & = & (n-m) & + & (m-p) & = & (n-m) & + & (m-p-1) & + & (1) \end{array}$$

where the second row above indicates the degrees of freedom. The single degree of freedom curvature test is a "quick and dirty" test, which can be applied whenever we run a 2-level factorial (full or a fraction) with center points. The null hypothesis is quite peculiar:

$$H_0 : \sum_{i=1}^{k} \beta_{jj} = 0$$

tested versus a two sided alternative. The test statistic is based on the single degree of freedom curvature sum of squares:

$$\text{SS}_{curvature} = \frac{n_f n_c (\overline{y}_f - \overline{y}_c)^2}{n_f + n_c}$$

where n_f and $\overline{y}_f$ are the number of factorial (corner) points on the design and the average response at those points, and n_c and $\overline{y}_c$ are the number of replicates at the center point ("number of center points") and the corresponding average response at the center. If the factorial portion of the experiment is *not* replicated, then $n - m = n_c + (m-1) - m = n_c - 1$, and the test statistic is

$$F_0 = \frac{\text{SS}_{curv.}/(1)}{\text{SS}_{PE}/(n_c - 1)}$$

which follows an F_{1,n_c-1} distribution if H_0 is true. We note how this test will not have much ability to detect curvature in saddle functions, where some $\beta'_{jj}s$ may be positive and some may be negative, thus the sum may be close to zero.

Appendix B
Analysis of Variance

Analysis of Variance, or ANOVA, is a technique developed mainly by R. Fisher in the 1920s to study statistical problems in genetics and improvement of crops. It corresponds to a generalization of the basic two-sample t-test to test for the equality of the means of two different populations. The generalization allows to test for the equality of means from 2 or more populations. In this appendix we outline the main ideas of ANOVA at an introductory level, sufficient to support the book sections where these ideas are utilized.

B.1 One Factor ANOVA, Fixed Effects

Suppose an experimenter is interested in comparing a different formulations of a new chemical product. The production of this chemical is in batches, and the response of interest is some property of the substance being produced, which can be measured in a metrology laboratory after a batch is prepared. The experiment she conducts consists of **replicating** each of the formulations by producing n batches of each formulation and taking a measurement of each batch. However, the way the experiment is conducted is in a **completely randomized** way. By this we mean that the order in which the an tests were conducted is at random. It also implies that the raw materials used in each trial are selected at random, not following any given sequence[1]. The experimental data can be arranged as in Table B.1. As shown on the table, let y_{ij} denote the observed response measured from the the jth replicate of the ith formulation. Note how the different formulations in our example correspond to different "levels" of the factor "formulation", so we are talking, in general, of a one factor experiment where the factor is varied across a different levels or **treatments**, a terminology that has its origin in agriculture and medical applications of ANOVA.

[1] We discuss randomization in Chapter 3.

Table B.1. Typical data collected from a one-way (one factor) experiment

Level of the factor (treatment)	**Replication number:**			
1	y_{11}	y_{12}	$\cdots$	y_{1n}
2	y_{21}	y_{22}	$\cdots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}

A model useful to model the response of such experiment is the **effects model**

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases} \tag{B.1}$$

where the $\varepsilon_{ij} \sim N(0, \sigma^2)$ are uncorrelated, and since they are assumed also normal, they are actually assumed independent[2]. What this model expresses is that the effect of varying the formulation can be represented by a constant (τ_i) that depends on the formulation. The other constant on the right hand side (μ) is the overall mean response, and the ε_{ij} are random errors we cannot observe directly and that cannot be attributed to anything else in concrete but to experimental error. This is a **fixed effects** model because the effect of the factor "formulation" is modeled as a constant. An alternative formulation is to let τ_i be a random variable which would result in a different model, called a **random effects model**. For the time being, we continue our discussion of the fixed effects model.

Once the effects model is postulated, the hypothesis of interest for the experimenter in this setting can only be whether the formulations (i.e., the different factor levels) make a difference or not (i.e., whether they have a significant *effect* or not). The hypothesis is then:

$$H_o: \ \tau_1 = \tau_2 = \ldots = \tau_a = 0 \tag{B.2}$$

or, if we define $\mu_i = \mu + \tau_i$ to be the mean response when the experiment was run at the ith level,

$$H_o: \ \mu_1 = \mu_2 = \ldots = \mu_a = \mu. \tag{B.3}$$

Written as (B.2), the hypothesis says that there are no differences from formulation to formulation, i.e., the effect of all treatments is the same and equal to zero. This is reflected if the hypothesis is written as in (B.3): if there are no treatment effects ($\tau_i = 0$ for all i) then this implies that the mean response will be the same regardless of what treatment (formulation) we choose.

A common interpretation of the treatment effects τ_i is that they are deviations from the overall mean. That is, if

$$\mu = \frac{\sum_{i=1}^{a} \mu_i}{a} \tag{B.4}$$

[2]Note that the variance is assumed to be constant for all observations regardless of the treatment and equal to σ^2.

then we must have that $\sum_{i=1}^{a} \tau_i = 0$, and therefore, each τ_i models *how much each treatment (formulation) changes on average the response from the overall mean*. Note how (B.4) imposes a constraint in the values of the treatment effects; if some are negative this equation implies that some other effects will be positive. As will be noted shortly, this is not the only interpretation possible, and different "constraints" on the parameters can be introduced in the model to allow their unique estimation, which will result in different interpretations.

To conduct the analysis of variance, most authors introduce the following ink-saving "dot notation":

$$y_{i\bullet} = \sum_{j=1}^{n} y_{ij}, \;\; \overline{y}_{i\bullet} = \frac{y_{i\bullet}}{n}, \quad i = 1, 2, \ldots, a$$

and

$$y_{\bullet\bullet} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}, \;\; \overline{y}_{\bullet\bullet} = \frac{y_{\bullet\bullet}}{N}$$

where $N = an$ denotes the total number of experiments. Thus, a dot in the place of a subscript replaces the summation operator over that subscript. When used with an overline, the dot subscript replaces the operation of averaging over that subscript.

With this notation, the Analysis of Variance consists in partitioning[3] the total variability in the observed experimental data with the goal of trying to attribute the ways in which the data can vary due to different causes or **sources**. In the simple one-way experiment we have discussed, the potential sources of variability in the data are either because we are changing the levels of the factor or because of some other unknown reason which we will call "error". The total variability in the data is measured by the total sum of squares[4]:

$$\text{SS}_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{\bullet\bullet})^2.$$

We want to partition this sum in:

$$\text{SS}_{total} = \text{SS}_{factor} + \text{SS}_{error}.$$

To do this, add and subtract $\overline{y}_{i\bullet}$ inside the square of SS_{total}, rearrange, and compute all products inside the square:

$$\begin{aligned} \text{SS}_{total} &= \sum_{i=1}^{a} \sum_{j=1}^{n} ((\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}) + (y_{ij} - \overline{y}_{i\bullet}))^2 = \\ &= n \sum_{i=1}^{a} (\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i\bullet})^2 \\ &= \text{SS}_{factor} + \text{SS}_{error} \end{aligned}$$

[3] "To partition" is precisely the meaning of the word "analysis".

[4] This is called the *corrected* sum of squares because it measures variability of the observed data around the average of the data, so we are "correcting for the mean".

where the final expression follows because the crossproduct is zero. This analysis gives a strategy for testing the null hypothesis (B.2) and determines whether or not there is a significant effect due to the factor. We should compare the variability we can attribute to changing the factor levels with the experimental error, and if the former is relatively large with respect to the latter, we should conclude the factor has a significant effect (or we say it "is significant"). In such case one would reject the null hypothesis, which indicates the factor has no effect. In order to do the corresponding comparison, the sum of squares being compared, which are essentially measures of variability, need to be on the same scale, because otherwise we would be comparing sums with very different number of terms. The degrees of freedom of each sum serve to scale each sum.

The **degrees of freedom** v of a sum of squares equals the constant such that $E[\text{SS}/v] = \sigma^2$, that is, it is the scaling constant that would make the sum of squares an unbiased estimator of the experimental error variance. In practice, the degrees of freedom are the number of quantities that are free to vary in a set of numbers for which we know one or more constraints. That is, suppose we have the set $S = \{y_1 = 5, y_2 = 8, y_3\}$ with y_3 unknown but we know that $\overline{y} = 5$. Can y_3 be any number we like? The answer is *no* because the average *constrains* the third element of S to be equal to 2, in such a way that the 3 numbers average 5. Thus, if the average is given, we say that set S has 2 degrees of freedom. If the average is not given we would say that S has 3 degrees of freedom.

The number of degrees of freedom is usually equal to the number of observations in a set minus the number of parameters that have been estimated (in the example above, one parameter was estimated, the mean). Applied to sum of squares, the number of degrees of freedom can usually be obtained by inspection: it is the number of "data elements" in the sum that can vary freely if the other quantities (usually averages) in the sum are given. Thus, we see that SS_{total} has $N - 1$ degrees of freedom because the last of the y_{ij}'s is constrained by the overall mean; SS_{factor} has $a - 1$ degrees of freedom, because the last one of the "data elements" in the sum, the *row averages* $\overline{y}_{i\bullet}$, is constrained by the overall average $\overline{y}_{\bullet\bullet}$ (i.e., the average of the averages is fixed). Finally, the sum of squares error has $an - a = a(n - 1)$ degrees of freedom because for each row average we lose one degree of freedom in the y_{ij}'s (we have an "data elements" and a constraints). Note that

$$N - 1 = an - 1 = (a - 1) + (an - a)$$

so the degrees of freedom of the sums of squares add up as well.

The information computed in the analysis of variance is usually summarized in books and by software according to an "ANOVA table" like that shown on Table B.2.

It is quite pedagogical to consider the simplest case when only $a = 2$ treatments are of interest (i.e., two formulations; we will use the generic term "treatments" from now on). If $a = 2$, the ANOVA table and the corresponding F test reduce to the familiar 2 sample t-test where we test $H_o : \mu_1 = \mu_2$ against a two sided alternative. Recall that in such case the test statistic is

$$t_0 = \frac{\overline{y}_{1\bullet} - \overline{y}_{2\bullet}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Table B.2. Analysis of Variance Table for a Single Factor Experiment, Fixed Effects Model

Source of variation	Sum of Squares	d.o.f.	Mean Squares	F_0
Factor (between treatments)	SS_{factor}	$a-1$	MS_{factor}	$F_0 = \frac{\text{MS}_{factor}}{\text{MS}_{error}}$
Error (within treatments)	SS_{error}	$a(n-1)$	MS_{error}	
Total	SS_{total}	$N-1$		

where S_p^2 is the "pooled" variance estimator that results from combining the sample variance estimators from each population, S_1^2 and S_2^2, using the corresponding degrees of freedom in each sample:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

If the null hypothesis is true, $t_0 \sim t_{n_1+n_2-2}$. An ANOVA test for $a=2$ should reduce to this case. Let us see how this is true.

From our earlier discussion, we have that $y_{ij} \sim N(\mu_i = \mu + \tau_i, \sigma^2)$. Let us assume for simplicity that $n_1 = n_2 = n$.

Suppose first H_o is *false*; then the τ_i's are not all zero, and the treatment means μ_i are not all equal to each other. Consider a sample of n observations taken at random from the (hypothetical) population corresponding to all observed responses that can possibly be obtained from when treatment one is applied to the process. This is actually what we get when we replicate the experiment n times and randomize the order of the runs and the assignment of treatments to the experimental units utilized. Since $a=2$, there are only two treatment effects, τ_1 and τ_2. Following the convention given by the constraint (B.4), we must then have that $\tau_1 = -\tau_2$.

For the case $a=2$ when $\mu_1 \neq \mu_2$, we see that we have "separation" between the two populations, i.e., the first sample, obtained from the first formulation, is centered around $\overline{y}_{1\bullet}$ and this will differ, on average, from where the second sample, obtained from the second treatment, is centered, namely, at $\overline{y}_{2\bullet}$. If we look at the SS_{error} term in this case for $i=1$ (say), we see that the squared deviations are distances from each data point y_{ij} from the corresponding average of that treatment, $\overline{y}_{i\bullet}$. Confining ourselves only to this first treatment ($i=1$), we have that if we divide the sum of squares by $n-1$, we get the sample variance of the first sample from it (what we use in the 2-sample t test), since

$$S_1^2 = \frac{\sum_{j=1}^{n}(y_{1j} - \overline{y}_{1\bullet})^2}{n-1}.$$

Similarly, looking at the SS_{error} term for the second treatment ($i=2$), we get, after dividing by $n-1$,

$$S_2^2 = \frac{\sum_{j=1}^{n}(y_{2j} - \overline{y}_{2\bullet})^2}{n-1}.$$

The sum $\sum_{i=1}^{2}$ in SS_{error} actually combines or "pools" these two sample variances, giving, in effect, the estimator used in the t-test:

$$S_p^2 = \frac{(n-1)S_1^2 + (n-1)S_2^2}{(n-1)+(n-1)} = \text{MS}_{error} = \widehat{\sigma}^2.$$

This estimator of σ^2, the experimental error variance, is an unbiased estimator no matter whether H_o is true or not.

Suppose now that H_o is *true* (and $a = 2$). In this case, $\mu_1 = \mu_2 = \mu$ and all treatments means $\overline{y}_{i\bullet}$ will appear to be generated from a distribution with mean μ and variance σ^2/n, that is,

$$\overline{y}_{i\bullet} \sim N(\mu, \sigma^2/n).$$

Looking at the SS_{factor} term in the ANOVA decomposition of the total sum of squares, we see that we can estimate $\sigma_{\overline{y}}^2 = \sigma^2/n$ from it by dividing by $a-1$:

$$\frac{\sum_{i=1}^{a}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}{a-1} = \widehat{\sigma}_{\overline{y}}^2 = \widehat{\sigma}^2/n.$$

Multiplying by n we get

$$\frac{n\ \sum_{i=1}^{a}(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2}{a-1} = \frac{\text{SS}_{factor}}{a-1} = \widehat{\sigma}^2.$$

Thus, *if H_o is true*, we get a second unbiased estimator of σ^2, this time from SS_{factor}. However, if H_o is *false*, then the treatment means are "separated" and the distances

$$\overline{y}_{1\bullet} - \overline{y}_{\bullet\bullet} \quad \text{and} \quad \overline{y}_{2\bullet} - \overline{y}_{\bullet\bullet}$$

which are squared in SS_{factor}, will *over*estimate σ^2.

Therefore, the main idea of ANOVA is that:

- when H_o is true, both MS_{factor} and MS_{error} are unbiased estimators of σ^2;
- when H_o is false, MS_{error} is again an unbiased estimator of σ^2 but MS_{factor} will be an overestimate of σ^2 on average

The ANOVA concludes by testing for the equality of these variance estimators by means of the F statistic:

$$F_0 = \frac{\text{MS}_{factor}}{\text{MS}_{error}}$$

which is distributed as an $F_{a-1,\ N-a}$ distribution provided H_o is true. This test has a right tail rejection area, since, if H_o is false, the numerator is expected to be larger than the numerator.

In the case when $a = 2$, if H_o is true we have that $F_0 \sim F_{1,\ 2n-2}$ and since $t_0^2 \sim F_{1,\ 2n-2}$ (from a well-known relation between the t and the F distributions) we see that the two sample t test is indeed a particular case of the one way ANOVA when we only have two treatments.

B.1.1 A More Formal Justification of ANOVA

The previous intuitive argument requires we explain two aspects of ANOVA in more detail:

1 Why is an F statistic used?

2 Why is $E[\text{MS}_{factor}] > \sigma^2$ if H_o is false and equal to σ^2 if H_o is true?

Let us answer the first question. If H_0 is true, it can be seen, from basic properties of the χ^2 distribution, that

$$\frac{\text{SS}_{factor}}{\sigma^2} \sim \chi^2_{a-1}.$$

Similarly, but independent of whether H_o is true or false,

$$\frac{\text{SS}_{error}}{\sigma^2} \sim \chi^2_{N-a}.$$

It is well known that ratios of *independent* χ^2 random variables follow an F distribution. Do these statistics follow independent χ^2 distributions? The answer is positive and is an application of a theorem by Cochran, a form which we state without proof next. It makes use of the fact that sum of independent squared standard normals follows a χ^2 distribution.

THEOREM B.1 *(Cochran's theorem). Let $Z_i \sim N(0,1)$ i.i.d. for $i = 1, 2, \ldots, v$ so that*

$$\sum_{i=1}^{v} Z_i^2 \sim \chi^2_v.$$

If

$$\sum_{i=1}^{v} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s \quad (s \leq v)$$

then each of the $Q_i \in \{Q_1, Q_2, \ldots, Q_s\}$ are independent $\chi^2_{v_i}$ random variables if and only if

$$\sum_{i=1}^{s} v_i = v.$$

Thus, if we have scaled sums of squares of normally distributed data that add up with corresponding degrees of freedom that adds up too, then the scaled sum of squares follow independent chi squared distributions. Applied to a one-factor ANOVA, we have that

$$Q_1 = \frac{\text{SS}_{factor}}{\sigma^2}, \quad v_1 = a - 1$$

and

$$Q_2 = \frac{\text{SS}_{error}}{\sigma^2}, \quad v_2 = N - a.$$

Since $v = N - 1 = v_1 + v_2$ are the degrees of freedom of

$$\frac{\text{SS}_{total}}{\sigma^2} = \sum_{i=1}^{N} Z_i^2 = Q_1 + Q_2$$

then $Q_1 = \frac{\text{SS}_{factor}}{\sigma^2}$ and $Q_2 = \frac{\text{SS}_{error}}{\sigma^2}$ are *independent* chi-squares.

Therefore,

$$\frac{\text{MS}_{factor}}{\text{MS}_{error}} = \frac{\frac{\text{SS}_{factor}}{\sigma^2}/(a-1)}{\frac{\text{SS}_{error}}{\sigma^2}/(N-a)} \sim \frac{\chi^2_{a-1}/(a-1)}{\chi^2_{N-a}/(N-a)} = F_{a-1,N-a}.$$

Thus, we would reject H_0 if $F_0 > F_{\alpha,a-1,N-a}$. In the case when $a = 2$, since the square of a t_v random variable follows an $F_{1,v}$ distribution, we see that the ANOVA test includes the 2-sample t test as a particular case.

B.1.2 Expected Sums of Squares, Fixed Effects One Way ANOVA

In this section we answer formally to the second question posed in the previous section, namely: Why is $E[\text{MS}_{factor}] > \sigma^2$ if H_o is false and equal to σ^2 if H_o is true? To answer this we need to find $E[\text{MS}_{factor}]$ and $E[\text{MS}_{error}]$.

Recall that

$$\text{SS}_{total} = \text{SS}_{Factor} + \text{SS}_{Error}$$

or

$$\text{SS}_{total} = \sum\sum(y_{ij} - \overline{y}_{\bullet\bullet})^2 = n\sum(\overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet})^2 + \sum\sum(y_{ij} - \overline{y}_{i\bullet})^2$$

The total sum of squares is:

$$\text{SS}_{total} = \sum\sum y_{ij}^2 - 2\overline{y}_{\bullet\bullet}\sum\sum y_{ij} + \frac{1}{an}(\sum\sum y_{ij})^2 = \sum\sum y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{an}$$

The treatment SS is:

$$\text{SS}_{factor} = n\sum\overline{y}_{i\bullet}^2 - 2n\overline{y}_{\bullet\bullet}\sum\overline{y}_{i\bullet} + \frac{1}{an}(\sum\sum y_{ij})^2 = \frac{1}{n}\sum y_{i\bullet}^2 - \frac{y_{\bullet\bullet}^2}{an}$$

Therefore,

$$\text{SS}_{error} = \text{SS}_{total} - \text{SS}_{factor} = \sum\sum y_{ij}^2 - \frac{1}{n}\sum y_{i\bullet}^2$$

We want to find $E(\text{MS}_{factor}) = E\left(\frac{\text{SS}_{factor}}{a-1}\right)$ and $E(\text{MS}_{error}) = E\left(\frac{\text{SS}_{error}}{N-a}\right)$.
The model is $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ for $i = 1, \ldots, a; j = 1, 2, \ldots, n$, and $N = an$. Some useful facts are that:

$$E(\varepsilon_{ij}) = E(\varepsilon_{i\bullet}) = E(\varepsilon_{\bullet\bullet}) = 0$$

$$E(\varepsilon_{ij}^2) = \sigma^2 \quad E(\varepsilon_{i\bullet}^2) = n\sigma^2 \quad E(\varepsilon_{\bullet\bullet}^2) = an\sigma^2$$

Then

$$E(\text{SS}_{factor}) = \underbrace{E\left(\frac{1}{n}\sum y_{i\bullet}^2\right)}_{\mathbf{1}} - \underbrace{E\left(\frac{y_{\bullet\bullet}^2}{an}\right)}_{\mathbf{2}}$$

Now,

$$\mathbf{1} = \frac{1}{n}\sum_{i=1}^{a} E(n\mu + n\tau_i + \varepsilon_{i\bullet})^2 = \frac{1}{n}\sum_{i=1}^{a}[(n\mu)^2 + n^2\tau_i^2 + n\sigma^2] = an\mu^2 + n\sum\tau_i^2 + a\sigma^2$$

Also,

$$\mathbf{2} = \frac{1}{an}E[(an\mu + n\sum_{i=1}^{a}\tau_i + \varepsilon_{\bullet\bullet})^2] = an\mu^2 + \sigma^2$$

Therefore,

$$E(\text{SS}_{factor}) = \mathbf{1} - \mathbf{2} = \sigma^2(a-1) + n\sum\tau_i^2$$

and

$$E\left(\frac{\text{SS}_{factor}}{a-1}\right) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

which is unbiased **only if** H_o **true**.
Similarly,

$$E[\text{SS}_{error}] = \underbrace{E(\sum\sum y_{ij}^2)}_{\mathbf{3}} - \underbrace{E\left(\frac{1}{n}\sum y_{i\bullet}^2\right)}_{\mathbf{1}}$$

Here we have that

$$\mathbf{3} = E[\sum\sum(\mu + \tau_i + \varepsilon_{ij})^2] = an\mu^2 + n\sum\tau_i^2 + an\sigma^2$$

so

$$E[\text{SS}_{error}] = \mathbf{3} - \mathbf{1} = \sigma^2(N - a)$$

and

$$E[\text{MS}_{error}] = E\left(\frac{\text{SS}_{error}}{N-a}\right) = \sigma^2$$

which is **always unbiased** regardless of H_o being true or not.

B.1.3 Parameter Estimation in ANOVA Fixed Effects Models

To estimate the $a + 1$ parameters in the one-way fixed effects model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

we can try applying the least squares criterion. The sum of squared errors is

$$L = \sum_{i=1}^{a}\sum_{j=1}^{n}\varepsilon_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \mu - \tau_i)^2.$$

From $\frac{\partial L}{\partial \mu} = 0$ and $\frac{\partial L}{\partial \tau_i} = 0$ $(i = 1, 2, \ldots, a)$ we get the system of "normal" equations (see Appendix A):

$$\begin{array}{ccccccc}
N\mu & +n\tau_1 & +\ldots & +n\tau_a & = & y_{\bullet\bullet} \\
n\mu & +n\tau_1 & & & = & y_{1\bullet} \\
n\mu & & +n\tau_2 & & = & y_{2\bullet} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
n\mu & & & +n\tau_a & = & y_{a\bullet}
\end{array}$$

This would appear as a system of $a+1$ equations with $a+1$ unknown parameters, however, the last a equations add up to the first one, so we really have a linearly independent equations. In stark contrast with the least squares estimation of regressions models where the normal equations are **full rank**, the normal equations in effects models are **rank deficient** or "less than full rank". As a consequence, there is an infinite number of solutions to this system of equations.

One way around this problem is to add an independent equation to the system above[5]. One usual equation that is added is

$$\sum_{i=1}^{a}\tau_i = 0$$

[5]As will be mentioned later,a better approach preferred by most authors (see [140]) is to use a Moore-Penrose generalized inverse which gives a unique solution.

(which, as mentioned before, implies treatment effects are deviations from the overall mean μ). With this equation, the solution obtained is

$$\widehat{\mu} = \overline{y}_{\bullet\bullet}, \quad \widehat{\tau}_i = \overline{y}_{i\bullet} - \overline{y}_{\bullet\bullet}, i = 1, 2, \ldots, a.$$

A natural question to ask is if there is any notion of uniqueness in this solution method, since the equation that was added is arbitrary. The answer is that no matter which equations are added to the system of normal equations, there are certain functions of the parameters that can always be estimated uniquely [140], the so-called **estimable functions**, such as: $\tau_i - \tau_j$, and $\mu_i = \mu + \tau_i$. In particular, predictions (point estimates) of the response, $\widehat{y}_{ij} = \mu_i$ are also uniquely estimated.

The effects model and the corresponding normal equations can be written using matrix notation. Suppose, for illustration, that $a = 2$ and $n = 3$. Then the effects model (B.1) can be written in the form of a linear regression model (see Appendix A) as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

or, if $a = 2$ and $n = 3$,

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{pmatrix}$$

An alternative way of writing the model is

$$\boldsymbol{y} = (\mathbf{1}_a \otimes \mathbf{1}_n)\boldsymbol{\mu} + (\boldsymbol{I}_a \otimes \mathbf{1}_n)\boldsymbol{\tau} + \boldsymbol{\varepsilon} \tag{B.5}$$

where $\mathbf{1}_k$ is a $k \times 1$ vector of ones, $\boldsymbol{I}_k$ is a $k \times k$ identity matrix, and $\boldsymbol{A} \otimes \boldsymbol{B}$ denotes the direct or Kronecker product of two matrices. If $\boldsymbol{A}$ is $r \times c$ and $\boldsymbol{B}$ is $s \times d$, the $\boldsymbol{A} \otimes \boldsymbol{B}$ is $rs \times cd$ and is obtained by taking each entry in the $\boldsymbol{A}$ matrix and multiplying it by the complete matrix $\boldsymbol{B}$, i.e., $\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij}\boldsymbol{B}]$ (see Appendix C).

The sum of squared errors is

$$L = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and from $\partial L/\partial\boldsymbol{\beta} = \mathbf{0}$ we get the normal equations

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}.$$

The $\boldsymbol{X}'\boldsymbol{X}$ matrix is symmetric and has the following structure for a general balanced one way ANOVA effects model:

$$\begin{pmatrix} N & n & n & \ldots & n & n \\ n & 0 & n & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n & 0 & 0 & \ldots & 0 & n \end{pmatrix}.$$

Thus, the solution revolves around finding an inverse for $\boldsymbol{X}'\boldsymbol{X}$. Since $\boldsymbol{X}$ has less than full rank (the last a columns add up to the first one in the example above) this results in an $\boldsymbol{X}'\boldsymbol{X}$ matrix that is rank deficient. We could add a constraint as mentioned earlier, but a better approach is to use a **generalized inverse** [140] of $\boldsymbol{X}'\boldsymbol{X}$. A generalized inverse[6] of a matrix $\boldsymbol{A}$ is a matrix $\boldsymbol{G}$ such that $\boldsymbol{AGA} = \boldsymbol{A}$. The solution[7] found would be given by

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{G}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{G}$ is a generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$. Methods for finding a generalized inverse are described in [140]; see Appendix C for a brief review. There is no single generalized inverse for a matrix, unless a Moore-Penrose generalized inverse is used (see Appendix C). However, there are estimable functions that are uniquely estimated regardless of the generalized inverse used. In particular, it can be shown that $\widehat{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}'\boldsymbol{y}$ is unique (i.e., the predictions are invariant to the choice of $\boldsymbol{G}$).

B.2 Random Effects Models

Situations occur when the levels of a factor in an experiment can be thought of as a sample of levels taken from a population of possible levels. The quantity of interest will not be, as in the fixed effects case, the *average* effect that the specific levels tried in the experiment would have on the response, but rather the effect the population of levels would have on the *variance* of the response. This is because if other levels had been sampled, we would have obtained other effect in the response. Thus, it is more interesting to ask how much will these differences in the response would be for *all* the possible levels that could be sampled. We then say the factor has a **random** effect, as opposed to a fixed effect. Variance estimation of the effect is the main goal whenever an experiment has random effects[8].

Perhaps the best example of a factor having a random effect is batches of product. If some properties of a product are being measured, and the production took place in batches, information is collected from a random sample of batches taken from a large warehouse or simply from the hypothetical population of all batches that can ever be produced, then "batches" can be thought of as a random effect factor.

The data collected and the corresponding statistical model for the one-way random effects model looks identical as in the fixed effects model (see equation (B.1)):

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

Here, however, it is traditionally assumed that

$$\text{Var}(\tau) = \sigma_\tau^2, \quad \text{Cov}(\tau_i, \tau_{i'}) = 0, \;\; i \neq i',$$

[6]Also called a pseudoinverse or g-inverse by some authors, and related to the Moore-Penrose inverse (the definition given turns out to be first Penrose condition, see Appendix C and Searle's excellent book [140]).
[7]Searle [140] points out that we should not call this an estimator, since it is arbitrary.
[8]Experiments with a mix of fixed and random effects are called *mixed effects*, as the Split-Plot designs discussed in Chapter 9.

and furthermore,

$$\text{Cov}(\tau_i, \varepsilon_{i',j}) = 0 \quad \text{for all } \; i, i', j$$

and since, as before, $\varepsilon_{ij} = \sigma^2_\varepsilon$, we have that $\sigma^2_y = \sigma^2_\tau + \sigma^2_\varepsilon$. Estimating the **variance components** σ^2_α and σ^2_ε is the main goal in random effects models. The corresponding hypothesis being tested is:

$$H_o : \sigma^2_\tau = 0$$

which, if true, would imply there is no random effect due to the factor. This means that for the batch-production example, no matter which batches (levels) we chose at random, their effect on the response will be the same, and hence, the variability of the response will not be affected.

The matrix formulation of the model has the same appearance as (B.1) or (B.5), but the traditional assumptions are

$$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon \boldsymbol{I}_n$$

$$\text{Var}(\boldsymbol{\tau}) = \sigma^2_\tau \boldsymbol{I}_a$$

and $\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\tau}) = \mathbf{0}_{n \times a}$. With this, we note that the observations will not have all the same variance, since from (B.5):

$$\text{Var}(\boldsymbol{y}) = (\boldsymbol{I}_a \otimes \mathbf{1}_a)\sigma^2_\tau \boldsymbol{I}_a (\boldsymbol{I}_a \otimes \mathbf{1}_n)' + \sigma^2_\varepsilon \boldsymbol{I}_n.$$

The analysis of variance in the random effects model is based on the same partition of the total variability as in the fixed effects case:

$$\text{SS}_{total} = \text{SS}_{Factor} + \text{SS}_{Error}$$

where

$$\text{SS}_{total} = \sum\sum y^2_{ij} - \frac{Y^2_{\bullet\bullet}}{N}$$

$$\text{SS}_{factor} = \frac{1}{n}\sum y^2_{i\bullet} - \frac{y^2_{\bullet\bullet}}{an}$$

and

$$\text{SS}_{error} = \text{SS}_{total} - \text{SS}_{factor} = \sum\sum y^2_{ij} - \frac{1}{n}\sum y^2_{i\bullet}.$$

B.2.1 Derivation of Expected Mean Squares, One-way Random Effects Model

We want to find $E(\text{MS}_{factor}) = E\left(\frac{\text{SS}_{factor}}{a-1}\right)$ and $E(\text{MS}_{error}) = E\left(\frac{\text{SS}_{error}}{N-a}\right)$. Recall the model is $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ for $i = 1, 2, 3, \ldots, a$ (selected at random); $j = 1, 2, \ldots, n$, and $N = an$. There are some new assumptions, compared with the fixed effects case:

$$E(\tau_i) = E(\tau_\bullet) = 0$$

$$E(\tau_i^2) = \sigma^2_\tau \quad E(\tau^2_\bullet) = a\sigma^2_\tau$$

$$E(\tau_i \varepsilon_{ij}) = 0$$

There are also the same assumptions for the errors terms as in fixed effects case, namely:

$$E(\varepsilon_{ij}) = E(\varepsilon_{i\bullet}) = E(\varepsilon_{\bullet j} = E(\varepsilon_{\bullet\bullet}) = 0$$

$$E(\varepsilon_{ij}^2) = \sigma^2 \quad E(\varepsilon_{i\bullet}^2) = n\sigma^2 \quad E(\varepsilon_{\bullet\bullet}^2) = an\sigma^2$$

We then have that

$$E(\text{SS}_{factor}) = \underbrace{E\left(\frac{1}{n}\sum y_{i\bullet}^2\right)}_{\mathbf{1}} - \underbrace{E\left(\frac{y_{\bullet\bullet}^2}{an}\right)}_{\mathbf{2}}$$

Now, let

$$\mathbf{1} = \frac{1}{n}\sum_{i=1}^{a} E(n\mu + n\tau_i + \varepsilon_{i\bullet})^2 = an\mu^2 + an\sigma_\tau^2 + a\sigma^2$$

(compare to the fixed effects case: $an\mu^2 + n\sum \tau_i^2 + a\sigma^2$). Also,

$$\begin{aligned}\mathbf{2} = \frac{1}{an}E[(an\mu + n\sum_{i=1}^{a}\tau_i + \varepsilon_{\bullet\bullet})^2] &= an\mu^2 + \frac{n^2 a\sigma_\tau^2}{an} + \sigma^2 \\ &= an\mu^2 + n\sigma_\tau^2 + \sigma^2\end{aligned}$$

(compare to the fixed effects case: $an\mu^2 + \sigma^2$). Therefore,

$$E(\text{SS}_{factor}) = \mathbf{1} - \mathbf{2} = \sigma^2(a-1) + \sigma_\tau^2(N-n)$$

(compare to $\sigma^2(a-1) + n\sum \tau_i^2$), and

$$E\left(\frac{\text{SS}_{factor}}{a-1}\right) = \sigma^2 + n\sigma_\tau^2$$

this is unbiased only if H_o true (compare to $\sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$ for the fixed effects case).

Similarly, we have that for the sum of squares of the error term

$$E[\text{SS}_{error}] = \underbrace{E(\sum\sum y_{ij}^2)}_{\mathbf{3}} - \underbrace{E\left(\frac{1}{n}\sum y_{i\bullet}^2\right)}_{\mathbf{1}}$$

Here we have that

$$\mathbf{3} = E[\sum\sum(\mu + \tau_i + \varepsilon_{ij})^2] = an\mu^2 + an\sigma_\tau^2 + an\sigma^2$$

(compare to $an\mu^2 + n\sum \tau_i^2 + an\sigma^2$), so

$$E[\text{SS}_{error}] = \mathbf{3} - \mathbf{1} = \sigma^2(N-a)$$

and

$$E[\text{MS}_{error}] = E\left(\frac{\text{SS}_{error}}{N-a}\right) = \sigma^2$$

Thus, MS_{error} is always unbiased, same as in the fixed effects case. The ANOVA information is summarized in a table as in Table B.3.

Table B.3. ANOVA table, one factor random effects model

Source of variation	S.S.	dof	MS	E(MS)
Factor (between treatments)	$n\sum_{i=1}^{a}(\overline{y}_{i\bullet}-\overline{y}_{\bullet\bullet})^2$	$a-1$	MS_{factor}	$\sigma^2+n\sigma_\tau^2$
Error (within treatments)	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\overline{y}_{i\bullet})^2$	$N-a$	MS_{error}	σ^2
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\overline{y}_{\bullet\bullet})^2$	$N-1$		

Therefore, following arguments similar than in the fixed effects case, to test $H_o : \sigma_\tau^2 = 0$, we compute:

$$F_0 = \frac{\text{MS}_{factor}}{\text{MS}_{error}}$$

Note how this is the *same* test statistic as in the 1-factor, fixed effects case, although the hypothesis differ. However, for more than one factor, or for cases where we have a mix of random and fixed effects, *the test statistics will differ from those in the fixed effects model*.

To determine which test statistic is needed to test a hypothesis when some factors are random, we need to compute the expected mean squares, and from them determine which ratios allow us to test the hypotheses of interest. See [141] for more information.

B.2.2 ANOVA Method of Estimating the Variance Components

Another important objective besides testing $H_o : \sigma_\tau^2 = 0$ is to estimate the variance components. A simple method for doing so, proposed by Fisher himself, is known as the "ANOVA method" of estimating the variance components. It is based on using the expected mean squares and equating them to the *observed* mean squares. This gives a system of equations with all the variance components as unknowns which is then solved. For the one-factor random effects model we just showed that

$$E(\text{SS}_{factor}) = (a-1)(n\sigma_\tau^2+\sigma_\varepsilon^2)$$

and that

$$E(\text{SS}_{error}) = a(n-1)\sigma_\varepsilon^2.$$

After conducting an experiment, we substitute the observed sums of squares on the left hand sides above and solve for the variance components. This yields,

$$\widehat{\sigma}_\varepsilon^2 = \frac{\text{SS}_{error}}{a(n-1)} = \text{MS}_{error}$$

which is as we would have expected, and

$$\sigma_\tau^2 = \left(\frac{\text{SS}_{factor}}{a-1} - \sigma_\varepsilon^2\right)/n = \frac{\text{MS}_{factor}-\text{MS}_{error}}{n}.$$

These estimators can be shown to be unbiased, but the second estimator can be negative whenever $\text{MS}_{error} > \text{MS}_{factor}$, something that can or cannot happen in actual practice as it only depends on the data. Because of this problem, the preferred method of estimation is REML (Restricted Maximum Likelihood) or Bayes approaches, which avoid this situation. See [141] for more information about these alternative estimation methods for variance components and their properties.

Appendix C
Matrix Algebra and Optimization Results

The purpose of this appendix is to provide a summary of definitions and notation in both linear algebra and optimization that are used in different places in the text.

C.1 Matrices

C.1.1 Basic Definitions

A *matrix* is a rectangular array of numbers called its elements. In this book, we confine ourselves to the case where the elements are real numbers. Matrices are denoted by bold uppercase letters. Its elements are denoted by a lowercase letter equal to that used for the matrix and two subscripts, giving the row and column location of the element. Thus,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}.$$

Thus $\boldsymbol{A}$ is a matrix having m rows and n columns, or a $m \times n$ matrix, where m and n are the *dimensions* of the matrix. If $m = n$, the matrix is said to be *squared*, and if $m = n = 1$ the matrix is referred to as a *scalar*. Scalar variables are *not* denoted by bold letters. Matrices with only one row are called row *vectors* and matrices with only one column are called column vectors. Vectors will be denoted by bold *lower*case letters.

The *transpose* of a $m \times n$ matrix is the $n \times m$ matrix $\boldsymbol{A}' = \boldsymbol{C}$ such that its elements are $c_{ij} = a_{ji}$. A squared matrix is *symmetric* if $\boldsymbol{A} = \boldsymbol{A}'$. Note that $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$.

The sum of matrices is only defined for matrices of the same dimension. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two $m \times n$ matrices, their sum is the $m \times n$ matrix whose elements are the sum of the corresponding elements of $\boldsymbol{A}$ and $\boldsymbol{B}$.

The multiplication or product of two matrices, say a matrix $\boldsymbol{A}$ of dimension $m \times n$ and a matrix $\boldsymbol{B}$ of dimension $n \times p$ equals the $m \times p$ matrix $\boldsymbol{C}$ such that its (i, j) element equals to

$$c_{i,j} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Multiplication is only defined for matrices that are *conformable*, i.e., the number of columns in $\boldsymbol{A}$ must equal the number of rows in $\boldsymbol{B}$. Notice that in general, $\boldsymbol{AB} \neq \boldsymbol{BA}$. Also, note that if $\boldsymbol{v}$ is $m \times 1$, then $\boldsymbol{vv}'$ is conformable and gives a $m \times m$ matrix. Lastly, note that if $\boldsymbol{v}$ is $n \times 1$ and $\boldsymbol{A}$ is $n \times n$, then $\boldsymbol{v}'\boldsymbol{Av}$ is a scalar.

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two $n \times 1$ vectors. The *scalar* or *dot* product of two vectors is defined as

$$\boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

The *norm* of a vector $\boldsymbol{x}$ gives its length, is denoted $||\boldsymbol{x}||$ and equals $\sqrt{\boldsymbol{x}'\boldsymbol{x}}$. Two vectors are said to be *orthogonal* if their dot product is zero, i.e., if $\boldsymbol{x}'\boldsymbol{y} = 0$ which geometrically it means the angle between them is 90°. The angle θ between two n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by the expression:

$$cos\ \theta = \frac{\boldsymbol{x}'\boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||}.$$

A set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ is said to be *linearly independent* if we cannot find scalars $\alpha_1, \alpha_2, \ldots, \alpha_k$, not all zero, such that $\sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$. Otherwise, the set of vectors is said to be linearly dependent. The set of all vectors that are linear combinations of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ is the *span* of these vectors, or the subspace spanned by the set of vectors. In this context, an n-dimensional *space* of vectors is the set of all possible vectors with n elements. In this book we only deal with n-dimensional Euclidean space, denoted $\mathbb{R}^n$ or E^n. A *subspace* S of $\mathbb{R}^n$ is a subset of $\mathbb{R}^n$ that is closed under vector addition and scalar multiplication, i.e., if $\boldsymbol{a}$ and $\boldsymbol{b}$ are two vectors and λ and μ are scalars, then $\lambda\boldsymbol{a} + \mu\boldsymbol{b}$ is in S.

The *rank* of a matrix is equal to the number of linearly independent columns which is also equal to the number of linearly independent rows. A $m \times n$ matrix $\boldsymbol{A}$ is said to be *full rank* if $rank(\boldsymbol{A}) = \min(m, n)$. This minimum is the largest number of independent vectors in the matrix. If $rank(\boldsymbol{A}) = m$ we say the matrix has full row rank and if $rank(\boldsymbol{A}) = n$ we say the matrix has full column rank.

The *trace* of a square matrix $\boldsymbol{A}$, denoted $trace(\boldsymbol{A})$ or $tr(\boldsymbol{A})$ equals to the sum of its diagonal elements. Useful properties of traces are:

1 $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$;

2 $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$.

3 $tr(\boldsymbol{x}'\boldsymbol{Bx})$ where $\boldsymbol{B}$ is s square matrix equals to $tr(\boldsymbol{Bxx}')$.

Let $\boldsymbol{A}$ be $m \times n$ and $\boldsymbol{B}$ be $p \times q$. Their *Kroenecker* or *direct* product is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = [a_{ij}\boldsymbol{B}]$$

and is a $mp \times nq$ matrix. That is, to compute the direct product take each element of $\boldsymbol{A}$ and multiply it by the whole matrix $\boldsymbol{B}$. Interesting properties of this product are that $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$ (provided the inverses exist), $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$ and that the product has the distributive property. The direct product does not have the commutative property.

Rules for differentiation.

1 Let $y = \boldsymbol{b}'\boldsymbol{x}$ where $\boldsymbol{b}$ is a vector which is not a function of $\boldsymbol{x}$. Then

$$\frac{\partial y}{\partial \boldsymbol{x}} = \boldsymbol{b}.$$

2 Let $y = \boldsymbol{x}'\boldsymbol{x}$. Then

$$\frac{\partial y}{\partial \boldsymbol{x}} = 2\boldsymbol{x}.$$

3 Let $y = \boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$ where $\boldsymbol{B}$ is a square matrix. Then

$$\frac{\partial y}{\partial \boldsymbol{x}} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{B}'\boldsymbol{x}.$$

C.1.2 Special Matrices

We denote by $\boldsymbol{I}$ the $m \times m$ identity matrix such that for any other $m \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}$. $\boldsymbol{I}$ is a squared diagonal matrix (a matrix with non-zero elements only along the diagonal) with diagonal elements all equal to 1. When the dimension of the identity matrix is not clear from the context it is made explicit by a subscript, e.g., $\boldsymbol{I}_m$.

An *orthogonal matrix* is a matrix $\boldsymbol{A}$ such that $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$. In this case, all vector columns of $\boldsymbol{A}$ are mutually orthogonal.

Let $\boldsymbol{X}$ be a $n \times p$ matrix of rank p. Let S_c be the space spanned by the columns of $\boldsymbol{X}$. Finally, let $\boldsymbol{Y}$ be an $n \times 1$ vector. The symmetric $n \times n$ matrix $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is called a **projection matrix**[1] because when applied to any any vector $\boldsymbol{Y}$, i.e., $\boldsymbol{P}\boldsymbol{Y}$, it projects $\boldsymbol{Y}$ from $\mathbb{R}^n$ on to S_c, i.e., it expresses $\boldsymbol{Y}$ as a linear combination of the columns of $\boldsymbol{X}$. If $\boldsymbol{P}$ is a projection matrix, then $\boldsymbol{I} - \boldsymbol{P}$ is also a projection matrix, which when applied to any vector $\boldsymbol{Y}$ projects this vector onto the subspace *orthogonal* to span$(\boldsymbol{X})$.

An interesting property of projections matrices is that they are **idempotent**. A matrix is idempotent if $\boldsymbol{A}^2 = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$, i.e., they are unchanged when multiplied times themselves.

C.1.3 Determinants and Matrix Inverse

The determinant of a $n \times n$ square matrix $\boldsymbol{A}$, written either det$(\boldsymbol{A})$ or $|\boldsymbol{A}|$, *equals the volume of the n-dimensional (hyper) parallelepiped generated by the rows of $\boldsymbol{A}$*, provided the *edges* of the parallelepiped come from the rows of $\boldsymbol{A}$. It also equals the volume of the parallelepiped generated by the columns of $\boldsymbol{A}$, and the volume is the same as that generated by the rows.

Example. Suppose

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 8 \\ -5 & 6 & 3 \\ 7 & 4 & 3 \end{bmatrix}.$$

[1]This is the "Hat" matrix used in regression analysis. See Appendix A.

The parallelepiped generated by the three rows of this matrix is shown in Figure C.1. The volume equals 448, and $\det(\boldsymbol{A}) = -448$, which gives this volume (the negative sign is due to the "left handed" orientation of the edges; it implies $\boldsymbol{A}$ is not positive definite). ■

It should be pointed out that two very different matrices can give the same determinant. This is true even for sum of squares matrices. For example,

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2.5 & 0 \\ 0 & 2 \end{bmatrix}$$

both have a determinant of 5, yet the rows of the second matrix generate a rectangle (with orthogonal "edges") while the first one generates a rombus.

The D-optimality criteria for experimental design (see Chapter 5) is based on maximizing the determinant of a sum of squares matrix.

Computation of the determinant. The determinant of matrix $\boldsymbol{A}$ is given by

$$det(\boldsymbol{A}) = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}$$

where the scalar

$$A_{ij} = (-1)^{i+j} \det(\boldsymbol{M}_{ij})$$

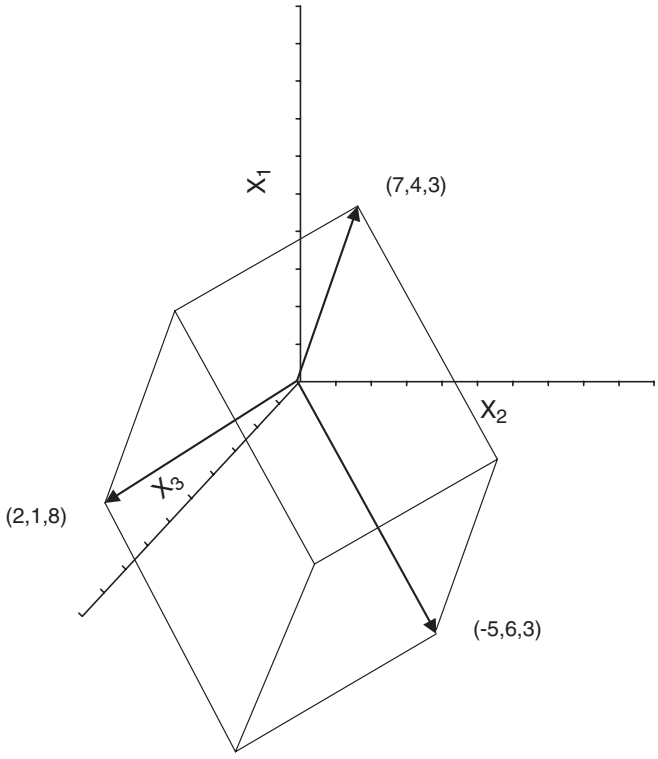


Figure C.1. Volume of the parallelepiped generated by the rows of matrix $\boldsymbol{A}$ in the example

is the *cofactor* that corresponds to the *minor* $\boldsymbol{M}_{ij}$ which is a matrix formed by deleting row i and column j from $\boldsymbol{A}$.

Matrix inverse

The inverse of an $n \times n$ square matrix $\boldsymbol{A}$ is defined as a matrix $\boldsymbol{A}^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$. It is computed from

$$\boldsymbol{A}^{-1} = \frac{\boldsymbol{A}_{cof}}{\det(\boldsymbol{A})}$$

where $\boldsymbol{A}_{cof}$ is the $n \times n$ matrix of cofactors of $\boldsymbol{A}$. This matrix may not exist; if $\det(\boldsymbol{A}) = 0$ the matrix is said to be *non-invertible*.

Some properties of the determinant and the inverse

1 $\det(\boldsymbol{I}) = 1$.

2 $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$.

3 $\det(c\boldsymbol{A}) = c^n \det(\boldsymbol{A})$, where c is a scalar (and $\boldsymbol{A}$ an $n \times n$ matrix).

4 $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$.

5 if $\boldsymbol{A}$ has a row or column equal to zero, $\det(\boldsymbol{A}) = 0$.

6 If $\boldsymbol{A}$ is invertible, then $\det(\boldsymbol{A}) \neq 0$. If $\boldsymbol{A}$ is singular, then $\det(\boldsymbol{A}) = 0$.

7 If $\det(\boldsymbol{A}) < 0$, then $\boldsymbol{A}$ is not positive definite.

8 If $\boldsymbol{A}$ and $\boldsymbol{B}$ are squared matrices of the same dimension, $|\boldsymbol{A}\boldsymbol{B}| = |\boldsymbol{A}||\boldsymbol{B}|$.

9 For an orthogonal matrix $\boldsymbol{A}$, since $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}$ we have that $\boldsymbol{A} = \boldsymbol{A}^{-1}$.

10 $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ provided the inverses exist.

The MATLAB command `det` returns the determinant of a square matrix $\boldsymbol{A}$.

C.1.4 Generalized Inverses and Solutions of Systems of Linear Equations

A **generalized inverse** or *g-inverse* of a matrix $\boldsymbol{A}$ is any matrix $\boldsymbol{G}$ that satisfies

$$\boldsymbol{A}\boldsymbol{G}\boldsymbol{A} = \boldsymbol{A}.$$

Note that a g-inverse as defined here may not be unique. Note also how this definition generalizes (and therefore, it includes) the usual concept of an inverse, since if a square matrix $\boldsymbol{A}$ is nonsingular, $\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}$ so $\boldsymbol{A}^{-1}$ is also a g-inverse (and in this case, it is unique).

A set of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is *consistent* if any linear relations among the rows of $\boldsymbol{A}$ also exist among the elements of $\boldsymbol{y}$. A system of equations can be solved if and only if the system is consistent.

If $\boldsymbol{Ax} = \boldsymbol{y}$ are consistent and have solution $\boldsymbol{x} = \boldsymbol{Gy}$, then $\boldsymbol{AGA} = \boldsymbol{A}$. Conversely, if $\boldsymbol{G}$ is such that $\boldsymbol{AGA} = \boldsymbol{A}$, then $\boldsymbol{Ax} = \boldsymbol{y}$ is consistent with solution $\boldsymbol{x} = \boldsymbol{Gy}$.

The *Moore-Penrose* g-inverse satisfies more conditions that make it unique. It satisfies:

$$\boldsymbol{AGA} = \boldsymbol{A}, \quad \boldsymbol{GAG} = \boldsymbol{G}, \quad (\boldsymbol{GA})' = \boldsymbol{GA}, \text{ and } (\boldsymbol{AG})' = \boldsymbol{AG}.$$

G-inverses of symmetric matrices. In linear models, we are primarily concerned with the $\boldsymbol{X}'\boldsymbol{X}$ matrix since we need to solve $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$. In *effects models* we do not have a full rank $\boldsymbol{X}$ matrix, and therefore $\boldsymbol{X}'\boldsymbol{X}$ is singular[2]. Hence we need to find a g-inverse of $\boldsymbol{X}'\boldsymbol{X}$. An important property of the g-inverse of $\boldsymbol{X}'\boldsymbol{X}$ is that if $\boldsymbol{G}$ is such g-inverse,

$$\boldsymbol{XGX}'$$

is *invariant* to the choice of g-inverse $\boldsymbol{G}$. This means that in a linear model,

$$\widehat{E[y]} = \widehat{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{XGX}'\boldsymbol{y}$$

is *estimable*, i.e., is uniquely estimated no matter our choice of $\boldsymbol{G}$ (see Appendix B).

Finding a G-inverse of a symmetric matrix

We wish to find a g-inverse of the squared, symmetric matrix $\boldsymbol{A}$. Let $\boldsymbol{B}$ and $\boldsymbol{C}$ be two orthogonal matrices that diagonalize $\boldsymbol{A}$, namely,

$$\boldsymbol{BAC} = \boldsymbol{\Delta}$$

where $\boldsymbol{\Delta}$ is diagonal. Then $\boldsymbol{G} = \boldsymbol{C}\boldsymbol{\Delta}^{-1}\boldsymbol{B}$ is a g-inverse of $\boldsymbol{A}$. If we set $\boldsymbol{B} = \boldsymbol{E}^{-1}$ and $\boldsymbol{C} = \boldsymbol{E}$, where $\boldsymbol{E}$ is the matrix of eigenvalues of $\boldsymbol{A}$, then $\boldsymbol{\Delta}$ is the diagonal matrix that has the eigenvalues of $\boldsymbol{A}$ along the diagonal (see C1.5 below). Thus,

$$\boldsymbol{G} = \boldsymbol{E}\boldsymbol{\Delta}^{-1}\boldsymbol{E}^{-1}$$

gives the desired G-inverse.

Example. Find a g-inverse of the matrix

$$\boldsymbol{A} = \begin{bmatrix} 2 & 4 & 5 \\ 4 & 3 & 6 \\ 5 & 6 & 7 \end{bmatrix}.$$

We have that

$$\boldsymbol{E} = \begin{bmatrix} 0.5916 & 0.6667 & 0.4534 \\ -0.7807 & 0.3333 & 0.5286 \\ 0.2012 & -0.6667 & 0.7177 \end{bmatrix}$$

and

$$\boldsymbol{\Delta} = \begin{bmatrix} -1.5777 & 0 & 0 \\ 0 & -1.0000 & 0 \\ 0 & 0 & 14.5777 \end{bmatrix}.$$

[2]When $\boldsymbol{X}$ has full column rank, $\boldsymbol{X}'\boldsymbol{X}$ is invertible.

Thus

$$\boldsymbol{G} = \boldsymbol{E}\boldsymbol{\Delta}^{-1}\boldsymbol{E}^{-1} = \begin{bmatrix} -0.6522 & 0.0870 & 0.3913 \\ 0.0870 & -0.4783 & 0.3478 \\ 0.3913 & 0.3478 & -0.4348 \end{bmatrix}. \quad \blacksquare$$

These computations are easily done in Matlab. The `eig` command returns matrix $\boldsymbol{E}$ and $\boldsymbol{\Delta}$ using

```
[E,Delta]=eig(A).
```

This is equivalent to the shorter and direct command `pinv(A)` which returns the g-inverse of matrix $\boldsymbol{A}$ in one step.

C.1.5 Eigenvalues, Eigenvectors, and Quadratic Forms

Consider a square $n \times n$ matrix $\boldsymbol{A}$. A scalar λ and a vector $\boldsymbol{x}$ that satisfy the nonlinear equation

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$$

(or $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$) are called an *eigenvalue* and an *eigenvector* of $\boldsymbol{A}$, respectively. Thus, the *scalar* λ is able to "mimic" matrix $\boldsymbol{A}$, hence the (German) name "characteristic" or "proper" (*eigen*). For λ to be an eigenvalue, it is necessary and sufficient that $\boldsymbol{A} - \lambda\boldsymbol{I}$ be a singular matrix[3] which implies that we can obtain the eigenvalues from the *characteristic equation* $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$. This gives a polynomial of order n in λ whose (possibly repeated) n roots are the eigenvalues. To find the eigenvectors, solve $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$. If in addition, $\boldsymbol{A}$ is symmetric, then:

1 all eigenvalues are real numbers;

2 eigenvectors associated with distinct eigenvalues are orthogonal;

3 the eigenvectors of $\boldsymbol{A}$ span the n-dimensional Euclidean space $\mathbb{R}^n$.

Let $\boldsymbol{W}$ be an $n \times n$ orthogonal matrix, i.e., $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n]$ where the $\boldsymbol{w}_i$'s span the n-dimensional Euclidean space. Recall that for an orthogonal matrix, $\boldsymbol{W}' = \boldsymbol{W}^{-1}$. Then for an $n \times n$ matrix $\boldsymbol{A}$ we have that

$$\begin{aligned} \boldsymbol{W}'\boldsymbol{A}\boldsymbol{W} = \boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{W} &= \boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{W} = \boldsymbol{W}^{-1}[\boldsymbol{A}\boldsymbol{w}_1, \boldsymbol{A}\boldsymbol{w}_2, \ldots, \boldsymbol{A}\boldsymbol{w}_n] \\ &= \boldsymbol{W}^{-1}[\lambda_1\boldsymbol{w}_1, \lambda_2\boldsymbol{w}_2, \ldots, \lambda_n\boldsymbol{w}_n] \\ &= \boldsymbol{W}^{-1}\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}\boldsymbol{W}. \end{aligned}$$

[3]This is a consequence of the non-zero solutions we want for $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{x} = \boldsymbol{0}$, which means that $\boldsymbol{A} - \lambda\boldsymbol{I}$ has a nullspace with vectors other than zero, that is, $\boldsymbol{A} - \lambda\boldsymbol{I}$ has linearly dependent columns and this can only happen if it is singular. Thus, we find the values of λ that make their determinant zero. That is, if for any square matrix $\boldsymbol{B}$ we have $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$ for $\boldsymbol{x} \neq \boldsymbol{0}$, this implies matrix $\boldsymbol{B}$ is singular.

Thus we have a means to *diagonalize* a matrix $\boldsymbol{A}$. This changes the basis, or coordinates, of the space and re-expresses them in terms of the space spanned by the eigenvectors of $\boldsymbol{A}$.

If $\boldsymbol{A}$ is $n \times n$, the product $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is a scalar and receives the name *quadratic form*. $\boldsymbol{A}$ is *positive definite* if the quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ is *positive* for all nonzero vectors $\boldsymbol{x}$. The quadratic form is *positive semi-definite, negative definite* and *negative semi-definite* if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0, < 0$, or ≤ 0 for all $\boldsymbol{x}$, respectively.

The connection between whether a matrix is positive definite or not and the eigenvalues can be seen from the diagonalization process above. Let $\boldsymbol{y} = \boldsymbol{W}^{-1}\boldsymbol{x}$ so $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{y}$. Then, from $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ we get $\boldsymbol{y}'\boldsymbol{W}'\boldsymbol{A}\boldsymbol{W}\boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i$. Since the y_i's are arbitrary, $\boldsymbol{A}$ can only be positive definitive (or positive semidefinite) if all its eigenvalues are positive (or nonnegative). If there are mixed-sign eigenvalues, then $\boldsymbol{A}$ is *indefinite* and the associated paraboloid is a saddle function.

Properties of eigenvalues. In what follows, let λ_i denote the eigenvalues of a $n \times n$ symmetric real matrix $\boldsymbol{A}$.

1 $\sum_{i=1}^{n} \lambda_i = tr(\boldsymbol{A})$

2 $\prod_{i=1}^{n} \lambda_i = det(\boldsymbol{A})$. It follows that if a square matrix has zero eigenvalues, it is singular.

C.2 Optimality Conditions

In this section we follow Luenberger [91]. For a function f of a n-dimensional vector $\boldsymbol{x}$, denoted $f(\boldsymbol{x})$, the *gradient* is defined as

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_n}\right]'$$

provided the derivatives are continuous. If all second derivatives are continuous, the *Hessian* of f at $\boldsymbol{x}$ is defined by the $n \times n$ symmetric matrix

$$\boldsymbol{F}(\boldsymbol{x}) = \left[\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}\right].$$

A point $\boldsymbol{x}^*$ is a *local minimum point* of a function f if there exists an $\varepsilon > 0$ such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{x}$ within a certain distance ε of $\boldsymbol{x}^*$ (i.e., for $|\boldsymbol{x} - \boldsymbol{x}^*| < \varepsilon$.) If $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$ for all $\boldsymbol{x}$, $\boldsymbol{x} \neq \boldsymbol{x}^*$, then $\boldsymbol{x}^*$ is a *global minimum (point)* of f.

Given a subset Ω of $\mathbb{R}^n$, the vector $\boldsymbol{d}$ is said to be a *feasible direction* at $\boldsymbol{x}$ if there is some α such that $\boldsymbol{x} + \beta\boldsymbol{d} \in \Omega$ for all $\beta \in (0, \alpha)$.

THEOREM C.1 **First order optimality conditions.** *Let Ω be a subset of $\mathbb{R}^n$. If $\boldsymbol{x}^*$ is a local minimum point of f (which is defined on Ω), then for any n dimensional feasible direction vector $\boldsymbol{d}$ at $\boldsymbol{x}^*$ we have that $\nabla f(\boldsymbol{x}^*)'\boldsymbol{d} \geq 0$.*

Notice this is a *necessary condition*, i.e., $\boldsymbol{x}^*$ local minimum $\Rightarrow \nabla f(\boldsymbol{x}^*)'\boldsymbol{d} \geq 0$ if $\boldsymbol{d}$ is a feasible direction. This condition says that at a local minimum, the angle between a feasible direction vector and the gradient vector is less than 90°. It is important to understand that this does not

tell us anything about the reverse implication: such angle can be less than 90° at some point $\boldsymbol{x}$, yet $\boldsymbol{x}$ may not be a local minimum.

COROLLARY C.2 *Consider the unconstrained case (i.e., $\Omega = \mathbb{R}^n$). If $\boldsymbol{x}^*$ is a local minimum point of f, then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$, i.e., at a local minimum the gradient vanishes.*

This points out a potential problem of steepest descent methods that try to move the operating conditions of a process along the direction implied by the gradient: as we approach a local minimum, the gradient becomes smaller, and hence harder to estimate from noisy data. There is an increasing need to use second order information as the optimization process approaches an optimum point. A test for curvature followed by fitting a quadratic model (if needed) serves this purpose. However, note that traditional RSM does not use second order information to move the process in a direction of improvement, it uses such information only to locate a local optimum (see [72] for an approach where second-order searches are used in RSM; the method, however, has received little attention in the Engineering Statistics community).

Let $\boldsymbol{x}^*$ be a point satisfying the nonlinear equality and inequality constraints

$$\boldsymbol{h}(\boldsymbol{x}^*) = \mathbf{0}, \quad \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}.$$

Let J be the set of inequality constraints for which $g_i(\boldsymbol{x}^*) = 0$, i.e., J is the set of *active constraints.* Assume there are p inequality constraints overall. Then point $\boldsymbol{x}^*$ is said to be a *regular point* if the gradient vectors $\boldsymbol{\nabla} h_i(\boldsymbol{x}^*)$ (for all m equality constraints i) and the vectors $\boldsymbol{g}_j(\boldsymbol{x}^*) = 0$ for all $j \in J$ are linearly independent. With this definition we can define the **Karush Kuhn Tucker (KKT)** first order necessary conditions for optimality of a general nonlinear programming problem.

THEOREM C.3 **Karush Khun Tucker (KKT) first order necessary conditions.** *Let $\boldsymbol{x}^*$ be a local minimum point of the problem*

$$\begin{aligned} &\textit{minimize} && f(\boldsymbol{x}) && \text{(C.1)} \\ &\textit{subject to} && \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}, \quad \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}. && \text{(C.2)} \end{aligned}$$

Suppose further that $\boldsymbol{x}^*$ is a regular point. Then there is a p-dimensional vector $\boldsymbol{\lambda}$ and a m-dimensional vector $\boldsymbol{\mu}$ such that

$$\begin{aligned} \boldsymbol{\nabla} f(\boldsymbol{x}^*) + \boldsymbol{\lambda}' \boldsymbol{\nabla} \boldsymbol{h}(\boldsymbol{x}^*) + \boldsymbol{\mu}' \boldsymbol{\nabla} \boldsymbol{g}(\boldsymbol{x}^*) &= \mathbf{0} \\ \boldsymbol{\mu}' \boldsymbol{g}(\boldsymbol{x}) &= \mathbf{0} \\ \boldsymbol{\mu} &\geq \mathbf{0} \end{aligned}$$

The first equation can be rewritten as: $\boldsymbol{\nabla} l(\boldsymbol{x}^*) = \mathbf{0}$, the gradient of the *Lagrangian* is the zero vector. It means that *the gradient of the objective function at $\boldsymbol{x}^*$ can be expressed as a linear combination of the gradient of the active constraints at $\boldsymbol{x}^*$*. The second expression is the *complementarity slackness* condition, which says that both the Lagrange multiplier μ_i and the inequality constraint $g_i(\boldsymbol{x})$ cannot be both greater than zero. These conditions together give a

system of $n + m + p$ equations in $n + m + p$ unknowns that in principle can be solved. Note that $\mu_i \geq 0$ but the λ's are not restricted in sign. Again, these are all necessary conditions for a local minimum, i.e., if $\boldsymbol{x}^*$ is a local minimum, then these conditions are true. But more importantly, it follows that *if the KKT conditions are* not *true at a point $\boldsymbol{x}^*$, then the point cannot be a local minimum* (and hence, a minimum). Thus the KKT conditions can be used to "test" whether a point satisfies them or not; if a point does satisfies them, in a sense this increases it "chances" of being an optimal point (see Chapter 13 for an application of this idea in simulation optimization). However, the conditions may be true but $\boldsymbol{x}^*$ may be not a local minimum. It is also important that no assumptions are made on the functions f, $\boldsymbol{h}$ and $\boldsymbol{g}$, other than they have continuous derivatives. In particular, no convexity assumption is made. We now state the second order conditions. In what follows, $\boldsymbol{H}$ and $\boldsymbol{G}$ are the Hessians of the equality and inequality constraints, respectively.

THEOREM C.4 **Second order necessary conditions.** *Suppose f, $\boldsymbol{h}$, and $\boldsymbol{g}$ have all continuous second derivatives. If $\boldsymbol{x}^*$ is a regular point and a local minimum of problem (C.1-C.2), then there is a m-dimensional vector $\boldsymbol{\lambda}$ and an p-dimensional vector $\boldsymbol{\mu}$, with $\boldsymbol{\mu} \geq \boldsymbol{0}$ such that the first order conditions hold and in addition the Hessian matrix*

$$\boldsymbol{L}(\boldsymbol{x}^*) = \boldsymbol{F}(\boldsymbol{x}^*) + \boldsymbol{\lambda}'\boldsymbol{H}(\boldsymbol{x}^*) + \boldsymbol{\mu}'\boldsymbol{G}(\boldsymbol{x}^*)$$

is positive semidefinite on the subspace $M = \{\boldsymbol{y} : \boldsymbol{\nabla h}(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{\nabla} g_j(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}$ for all $j = \in J\}$, where $J = \{j : g_j(\boldsymbol{x}^*) = 0\}$, i.e., the space of the active constraints at $\boldsymbol{x}^*$.

The condition on the Hessian (positive definiteness on M) can be stated as $\boldsymbol{y}'\boldsymbol{L}(\boldsymbol{x}^*)\boldsymbol{y} \geq 0$ for all $\boldsymbol{y} \in M$. These conditions are necessary, and include the case when there are inequality constraints that are active but have a zero Lagrange multiplier (these are "degenerate" inequality constraints). *Sufficient second order conditions* can also be stated, i.e., conditions that if true guarantee that $\boldsymbol{x}^*$ is a local minimum. These exclude the case of degenerate inequality constraints, and hence the subpace M' below accounts for that.

THEOREM C.5 **Second order sufficient conditions.** *Suppose f, $\boldsymbol{h}$, and $\boldsymbol{g}$ have all continuous second derivatives. A point $\boldsymbol{x}^*$ is a local minimum point of problem (C.1–C.2) if all of the following holds:*

1. *$\boldsymbol{x}^*$ is a regular point;*
2. *the first order necessary conditions are true;*
3. *the Hessian matrix of the Lagrangian*

$$\boldsymbol{L}(\boldsymbol{x}^*) = \boldsymbol{F}(\boldsymbol{x}^*) + \boldsymbol{\lambda}'\boldsymbol{H}(\boldsymbol{x}^*) + \boldsymbol{\mu}'\boldsymbol{G}(\boldsymbol{x}^*)$$

is positive definite on the subspace

$$M' = \{\boldsymbol{y} : \boldsymbol{\nabla h}(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{\nabla} g_j(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0} \text{ for all } j \in J'\}$$

where $J' = \{j : g_j(\boldsymbol{x}^*) = 0, \mu_j > 0\}$.

Example. Illustration of optimality conditions in a mixture experiment. Suppose we wish to solve the following problem

$$\text{maximize} \quad 10 + x_1 + 2x_2 + x_3 + x_1x_2 + x_2x_3 + x_1x_3$$

subject to

$$x_1 + x_2 + x_3 = 1 \quad x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0.$$

We have that

$$\boldsymbol{\nabla} f(\boldsymbol{x}) = \begin{bmatrix} 1 + x_2 + x_3 \\ 2 + x_1 + x_3 \\ 1 + x_2 + x_1 \end{bmatrix}$$

so

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Also,

$$\boldsymbol{\nabla} h(\boldsymbol{x}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

and therefore $\boldsymbol{H}(\boldsymbol{x}) = \boldsymbol{0}_{3\times 3}$. Finally, we have that $\boldsymbol{\nabla g}(\boldsymbol{x}) = -\boldsymbol{I}_3$. Note how none of $\boldsymbol{F}$, $\boldsymbol{\nabla} h$, $\boldsymbol{H}$, or $\boldsymbol{\nabla g}$ are functions of $\boldsymbol{x}$, so they hold for any point $\boldsymbol{x}$.

If we neglect for a moment the inequality constraints, the first order necessary conditions for a maximum are $\boldsymbol{\nabla} l(\boldsymbol{x}) = \boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\lambda}'\boldsymbol{\nabla} h(\boldsymbol{x}) = \boldsymbol{0}$, or:

$$\begin{aligned} 1 + x_2 + x_3 - \lambda &= 0 \\ 2 + x_1 + x_3 - \lambda &= 0 \\ 1 + x_2 + x_1 - \lambda &= 0 \end{aligned}$$

together with the equality constraint $x_1 + x_2 + x_3 = 1$. Solving this system of equations yields the unique solution $\boldsymbol{x}^* = (0, 1, 0)'$, $\lambda = 2$. This gives $f(\boldsymbol{x}^*) = 12$ and $h(\boldsymbol{x}^*) = 1$. Note how this solution also satisfies the inequality constraints.

We can determine if this $\boldsymbol{x}^*$ is indeed a local maximum by checking the second order sufficient conditions (again, neglecting the inequality constraints). These indicate that the Hessian of the Lagrangian,

$$\boldsymbol{L}(\boldsymbol{x}^*) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

must be positive definite on the space $M = \{\boldsymbol{y} : \boldsymbol{\nabla} h(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}\}$ and since $\boldsymbol{\nabla} h(\boldsymbol{x}^*)'\boldsymbol{y} = y_1 + y_2 + y_3 = 0$, we have that

$$(y_1, y_2, y_3) \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = -(y_1^2 + y_2^2 + y_3^2).$$

Thus, $\boldsymbol{L}$ is negative definite on $\boldsymbol{M}$, and therefore $\boldsymbol{x}^* = (0, 1, 0)'$ is a local *maximum*. However, note that *any point* that satisfies the necessary conditions also satisfies the sufficient conditions (since they are not a function of $\boldsymbol{x}$). Therefore, in this problem, the necessary conditions are also sufficient. This is also true if we consider the inequality constraints, as shown next.

Now, if we consider in this problem the inequality constraints, the first order necessary conditions are:

$$\begin{aligned} 1 + x_2 + x_3 - \lambda - \mu_1 &= 0 \\ 2 + x_1 + x_3 - \lambda - \mu_2 &= 0 \\ 1 + x_2 + x_1 - \lambda - \mu_3 &= 0 \\ x_1 + x_2 + x_3 &= 1 \\ \mu_1 x_1 &= 0 \\ \mu_2 x_2 &= 0 \\ \mu_3 x_3 &= 0 \end{aligned}$$

and $\mu_i \geq 0$, $i = 1, 2, 3$. Solving the system of equalities above, we see that one of the solutions is $\boldsymbol{x}^* = (0, 1, 0)'$ with $\lambda = 2$, and $\mu_1 = \mu_2 = \mu_3 = 0$. Thus, although the inequalities $x_1 \geq 0$ and $x_3 \geq 0$ are active at $\boldsymbol{x}^*$, they are degenerate since their Lagrange multipliers are zero. Thus, we see that this solution $\boldsymbol{x}^*$ satisfies the first order conditions for the inequality constraint case.

Finally, we can check the second order sufficient conditions. In this case, $M = \{\boldsymbol{y} : \boldsymbol{\nabla h}(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}\}$, and this is also equal to $M' = \{\boldsymbol{y} : \boldsymbol{\nabla h}(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}, \boldsymbol{\nabla} g_j(\boldsymbol{x}^*)'\boldsymbol{y} = \boldsymbol{0}$ for all $j \in J'\}$ since the set $J' = \{j : g_j(\boldsymbol{x}^*)' = 0, \mu_j > 0\} = \varnothing$. Similarly as before, we find that $\boldsymbol{L}(\boldsymbol{x}^*)$ is negative definite on M (and M') and therefore $\boldsymbol{x}^*$ is at least a local maximum. Since the Hessian is not a function of $\boldsymbol{x}$, this means that the necessary conditions are also sufficient for any point $\boldsymbol{x}$ that satisfies them. ■

In summary, optimality conditions can only guarantee whether a point is a local minimum/maximum. Checking the sufficient conditions can be quite a difficult task in general problems. Therefore, practically all gradient-based optimization methods stop at a "Karush-Khun-Tucker" (KKT) point, i.e., at a point that simply satisfies the first order necessary conditions. But since these are only necessary, the optimization method can converge to points that are *not even local minima/maxima.* Thus, it is always a "hope" that a point that satisfies the necessary conditions is a minimum/maximum, and this is reasonable in as much as these are conditions that must be satisfied at a true local minimum/maximum, as the sufficient conditions include the necessary conditions as a subset. Thus, if a point does not satisfy the KKT conditions, it cannot be optimal. For this reason, numerical optimization algorithms for non-linear programming are usually started from a grid of starting points, in an attempt to detect the global

optimum or at least find a good local optimum. Latin hypersquares (see Chapter 14) of initial points are sometimes used for this purpose.

Evidently, the discussion in this section applies when we are dealing with a complicated, non-convex function. If the objective function and the constraints are convex, specialized algorithms exist that guarantee global optimality. See [5].

Appendix D
Some Probability Results Used in Bayesian Inference

This brief Appendix gives some results that are used in the book, mainly in Part V.

Change of Variable Theorem in Definite Integrals

Assume the composed function $f \circ g$ is defined, f is continuous and g has a continuous derivative on $[a, b]$. Then

$$\int_{g(a)}^{g(b)} f(g)dg = \int_a^b f(g(x))g'(x)dx.$$

Transformation of Random Variables Theorem

Let X be a continuous random variable with density $p(x)$ and assume $Y = u(X)$ is a one-to-one transformation from $A = \{x : p(x) > 0\}$ to $B = \{y : p(y) > 0\}$ with inverse transformation $x = u^{-1}(y) = w(y)$. If the derivative $d/dy\ w(y)$ is continuous and nonzero in B, the density of Y is given by:

$$p(y) = p(w(y)) \left| \frac{d\ w(y)}{dy} \right|.$$

Scaled Inverse χ^2 and Inverse Gamma Distributions

The Inv$-\chi^2(v_0, \sigma_0^2)$ (scaled inverse chi-squared) is the distribution of $\sigma_0^2 v_0^2/\chi_{v_0}^2$, i.e., it is the inverse of a usual χ^2 distribution with v_0 degrees of freedom that is scaled by the quantity $\sigma_0^2 v_0^2$, hence its name. Its density is

$$p(\theta) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}\ s^\nu\ \theta^{-(v/2+1)}\ e^{-\nu s^2/(2\theta)}, \quad \theta > 0$$

which has mean $E(\theta) = \frac{\nu}{\nu-2}s^2$, $(\nu > 2)$, Mode$(\theta) = \frac{\nu}{\nu+2}s^2$ and Var$(\theta) = \frac{2\nu^2 s^4}{(\nu-2)^2(\nu-4)}$, $(\nu > 4)$.

The conjugate prior for the variance of a normal is the Inverse Gamma distribution ($IG(\alpha, \beta)$), which density is

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\ \theta^{-(\alpha+1)}\ e^{-\beta/\theta}, \quad \theta > 0$$

with mean $E[\theta] = \frac{\beta}{\alpha-1}$, $(\alpha > 1)$, Mode$(\theta) = \frac{\beta}{\alpha+1}$, and Var$(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, (\alpha > 2)$.

The Scaled inverse χ^2 distribution is then a particular IG$(\nu/2, \nu s^2/2)$ distribution, hence it is the conjugate prior distribution for the normal variance.

Inverse Wishart Distribution

This is the conjugate prior distribution for the covariance matrix of a multivariate normal distribution. It is the multivariate generalization of the scaled inverse χ^2. If $\boldsymbol{W}$ is a $p \times p$ positive definite matrix, then the IW density is

$$p(\boldsymbol{W}) = \left(2^{\nu p/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} |\boldsymbol{S}|^{-\nu/2}|\boldsymbol{W}|^{(\nu-p-1)/2}\ e^{-\frac{1}{2}tr(\boldsymbol{S}^{-1}\boldsymbol{W})}$$

with mean equal to $E[\boldsymbol{W}] = \nu\boldsymbol{S}$.

(Scalar) Student t Density Function

$$p(t) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{v\pi}\sigma}\left(1 + \frac{1}{v}\left(\frac{t-\mu}{\sigma}\right)^2\right)^{-(v+1)/2}$$

where $E(t) = \mu = \text{mode}(t)$, Var$(t) = \frac{v}{v-2}\sigma^2$ $(v > 2)$.

Multivariate Student t Density Function

A $q \times 1$ random vector $\boldsymbol{t}$ is distributed as a (non-central) multivariate t if its density is

$$p(\boldsymbol{t}) = \frac{\Gamma((v+q)/2)}{\Gamma(v/2)v^{q/2}\pi^{q/2}}|\boldsymbol{\Sigma}|^{-1/2}\left(1 + \frac{1}{v}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right)^{-(v+q)/2}$$

where $E(\boldsymbol{t}) = \boldsymbol{\mu} = \text{mode}(\boldsymbol{t})$, Var$(\boldsymbol{t}) = \frac{v}{v-2}\boldsymbol{\Sigma}$ $(v > 2)$.

Matrix T Density

A $l_1 \times l_2$ random matrix $\boldsymbol{T}$ follows a (central) matrix T distribution with parameters $(v, \boldsymbol{P}, \boldsymbol{Q})$ if its density is

$$p(\boldsymbol{T}) = \frac{k(v, l_1, l_2)}{|\boldsymbol{P}|^{v/2}|\boldsymbol{Q}|^{l_2/2}}\frac{1}{|\boldsymbol{P}^{-1} + \boldsymbol{T}\boldsymbol{Q}^{-1}\boldsymbol{T}'|^{(v+l_2)/2}}$$

where $\boldsymbol{P}$ is $l_1 \times l_1$, and $\boldsymbol{Q}$ is $l_2 \times l_2$ $(\boldsymbol{Q} > 0)$, and $E(\boldsymbol{T}) = \boldsymbol{0} = \text{mode}(\boldsymbol{T})$, and Var$(vec(\boldsymbol{T}')) = \frac{v}{v-2}\boldsymbol{P}^{-1} \otimes \boldsymbol{\Sigma}$ (a $(l_1 \times l_2) \times (l_1 \times l_2)$ matrix).

References

[1] Aitken, A.C., *Determinants and Matrices*, Oliver and Boyd, 1958.

[2] Aleksandrov, A.D., Kolmogorov, A.N., and Lavrent'ev, M.A., *Mathematics, Its Content, Methods and Meaning*, NY: Dover, 1999.

[3] Arnold, S.F. *Mathematical Statistics*. Englewood Cliffs, NJ: Prentice Hall, 1990.

[4] Banks, J. Carson, J., Nelson, B.L., and Nicol, D., *Discrete-Event Simulation*, 4th. ed., NY: Prentice Hall, 2004.

[5] Bazaraa, M.S., Sherali, H.D., and Shetty, C.M., *Nonlinear Programming: Theory and Algorithms*, 2nd Ed., NY: Wiley, 1993.

[6] Becker, N.G., "Models for the response of a mixture", *Journal of the Royal Statistical Society, Series B (Methodological)*, 30, 1968.

[7] Berger, J.O., *Statistical Decision Theory and Bayesian Analysis*, 2nd Ed., New York: Springer-Verlag, 1985.

[8] Berliner, L.M., "Bayesian control in mixture models", *Technometrics*, 29, 4, pp. 455–460, 1987.

[9] Bettonvil, B., del Castillo, E., Kleijnen, J.P.C., "Statistical testing of optimality conditions in multiresponse simulation-based optimization," Technical paper, University of Tilburg Center for Econometrics Research, http://greywww.kub.nl:2080/greyfiles/center/2005/81.html, 2005.

[10] Bisgaard, S., "The design and analysis of $2^{k-p} \times 2^{q-r}$ Split Plot Experiments", *Journal of Quality Technology*, 32,1, pp. 39–56, 2000.

[11] Bisgaard, S., and Ankenman, B., "Standard errors for the eigenvalues in second-order response surface models", *Technometrics*, 38, pp. 238–246, 1996.

[12] Blum, J., "Multivariate stochastic approximation methods", *Ann. Math. Stat.* 25, pp. 737–744, 1954.

[13] Borkowski, J.J., and Lucas, J.M., "Designs of mixed resolution for process robustness studies", *Technometrics*, 39, pp. 63–70, 1997.

[14] Borror, C.M., Montgomery, D.C., and Myers, R.H., "Evaluation of statistical designs for experiments involving noise factors", *Journal of Quality Technology*, 34,1, pp. 54–70, 2002.

[15] Borror, C.M., and Montgomery, D.C., "Mixed resolution designs as alternatives to Taguchi inner/outer array designs for robust design problems", *Quality and Reliability Engineering International*, 16, 2, pp. 117–127, 2000.

[16] Box, G.E.P., "Multifactor designs of first order", *Biometrika*, 39, 1–2, pp. 49–57, 1952.

[17] Box, G.E.P., "Contribution to the discussion, Symposyoum on interval estimation", *Journal of the Royal Statistical Society, B*, 16, pp. 211–212, 1954.

[18] Box, G.E.P., "Signal to noise ratios, performance criteria, and transformations", *Technometrics*, 30, pp. 1–40, 1988.

[19] Box, G.E.P., and Behnken, D.W., "Some new three-level designs for the study of quantitative variables", *Technometrics*, 2, pp. 455–475, 1960.

[20] Box, G.E.P., and Draper, N.R., *Empirical Model-Building and Response Surfaces*. NY: John Wiley & Sons, 1987.

[21] Box, G.E.P., and Draper, N.R., "A basis for the selection of a response surface design", *Journal of the American Statistical Association*, 54, pp. 622–654, 1959.

[22] Box, G.E.P., and Draper, N.R., "The choice of a second order rotatable design", *Biometrika*, 50, pp. 335–352, 1963.

[23] Box, G.E.P., and Hunter, J.S., "A confidence region for the solution of a set of simulataneous equations with an application to experimental design", *Biometrika*, 41, pp. 190–199, 1954.

[24] Box, G.E.P., and Hunter, J.S., "Multifactor experimental designs for exploring response surfaces", *Ann. Math., Statist.*, 28, pp. 195–242, 1957.

[25] Box, G.E.P., Hunter, J.S., and Hunter, W.G., *Statistics for Experimenters*, 2nd ed. Wiley, NY, 2005.

[26] Box G.E.P., G.M. Jenkins, and Reinsel, G., *Time Series Analysis, Forecasting, and Control* 3rd. ed., Englewood Cliffs: Prentice Hall, 1994.

[27] Box, G.E.P., and Jones, S., "Designing Products That Are Robust To The Environment. *Report Series in Quality and Productivity*. CPQI, University of Wisconsin, Number 56, 1990.

[28] Box, G.E.P., and Tiao, G.C., *Bayesian Inference in Statistical Analysis*, Reading, Mass: Addison Wesley, 1973.

[29] Box, G.E.P., and Wilson, K.B. "On the experimental attainment of optimum conditions," *Journal of the Royal Statistical Society*, B13, 1–38, 1951.

[30] Cahya, S., *Sampling Properties of Optimal Operating Conditions of Single and Multiple Response Surface Systems*, PhD. Dissertation, Dept. of Industrial Engineering, Penn State University, 2001.

[31] Cahya, S., del Castillo, E., and Peterson, J.J., "Computation of confidence regions for optimal factor levels in constrained response surface problems," *Journal of Computational and Graphical Statistics*, 13(2), pp. 1–20, 2004.

[32] Cook, R.D., "Influential observations in linear regression", *Journal of the American Statistical Association*, 74, pp. 169–174, 1977.

[33] Cook, R.D., and Nachtsheim, C.J., "A comparison of algorithms for constructing exact D-optimal designs", *Technometrics*, 22, 3, pp. 315–324, 1980.

[34] Cornell, J.A., *Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data*, 2nd ed., NY: Wiley, 1990.

[35] Cressie, N.A.C., *Statistics for Spatial Data*, NY: Wiley, 1993.

[36] Czitrom, V., and Spagon, P.D., *Statistical Case Studies for Industrial Process Improvement*, American Statistical Association and the Society for Industrial and Applied Mathematics, 1997.

[37] del Castillo, E., "Stopping rules for steepest ascent in experimental optimization," *Communications in Statistics, Simulation and Computation*, 26(4), 1599–1615, 1997.

[38] del Castillo, E., "Multiresponse optimization via constrained confidence regions," *Journal of Quality Technology*, 28, 1, pp. 61–70, 1996.

[39] del Castillo, E., *Statistical Process Adjustment for Quality Control*, NY: Wiley, 2002.

[40] del Castillo, E., M.J. Alvarez, L. Izarbe, and E. Viles, "A new design criterion for robust parameter experiments", to appear in *Journal of Quality Technology*, 2007.

[41] del Castillo, E., and Cahya, S., "A tool for computing confidence regions on the stationary point of a response surface", *The American Statistician*, 55, 4, pp. 358–365, 2001.

[42] del Castillo, E., Fan, S.K., and Semple, J., "Optimization of dual response systems: A Comprehensive Procedure for Degenerate and Nondegenerate Problems," *European Journal of Operational Research*, 112, pp. 174–186, 1999.

[43] del Castillo, E., Fan, S.K., and Semple, J., "The Computation of Global Optima in Dual Response Systems," *Journal of Quality Technology*, 29, 3, pp. 347–353, 1997.

[44] Delozier, M.R., "Application of Response Surfaces in Evaluating Tool Performance in Metalcutting", Private communication, 1999.

[45] Derringer, G.C., and Suich, R., "Simultaneous optimization of several response variables," *Journal of Quality Technology*, 12, pp. 214–219, 1980.

[46] Diaz-Garcia, J.A., and Ramos-Quiroga, R., "An approach to optimization in response surface methodology", *Communications in Statistics, Theory and Methods*, 30, 5, pp. 827–835, 2001.

[47] Diaz-Garcia, J.A., Ramos-Quiroga, R., and Cabrera-Vicencio, E., "Stochastic programming methods in the response surface methodology", *Computational Statistics and Data Analysis*, 49, pp. 837–849, 2005.

[48] Draper, N.R., "Ridge analysis of response surfaces", *Technometrics*, 5, pp. 469–479, 1963.

[49] Draper, N.R., and Lin, D.K.J., "Small response surface designs", *Technometrics*, 32, pp. 187–194, 1990.

[50] DuMouchel, W., and Jones, B., "A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model", *Technometrics*, 36, 1, pp. 37–47, 1994.

[51] Dvoretzky, A., "On stochastic approximation". *Proceedings of the 3rd Berkeley Symposium on Mathematical Stats.*, 1955.

[52] Dykstra, O., "The augmentation of experimental data to maximize $|\boldsymbol{X}'\boldsymbol{X}|$." *Technometrics*, 13, pp. 682–688, 1971.

[53] Efron, B., and Tibshirani, R.J., *An Introduction to the Bootstrap*, NY: Chapman & Hall/CRC Press, 1994.

[54] Fedorov, V.V., *Theory of Optimal Experiments*, NY: Academic Press, 1972.

[55] Fisher, R.A., *The Design of Experiments*, 5th ed., London: Hafner, 1949.

[56] Frey, D.D., and Wang, H., "Adaptive one-factor-at-a-time experimentation and expected value of improvement", *Technometrics*, 48, 3, pp. 418–431, 2006.

[57] Frisbee, S.E., and McGinity, J.W., "Influence of nonionic surfactants on the physical and chemical properties of a biodegradable Pseudolatex," *Eur. J. Pharm Biopharm.*, 40(6), 355–363, 1994.

[58] Galil, S., and Kiefer, J., "Comparison of Box-Draper and D-optimum designs for experiments with mixtures", *Technometrics*, 19, 4, pp. 441–444, 1977.

[59] Ganju, J., and Lucas, J.M., "Bias in test statistics when restrictions in randomization are caused by factors", *Communications in Statistics, Theory and Methods*, 26, pp. 47–63, 1997.

[60] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., *Bayesian Data Analysis*, 2nd Ed. Boca Raton Fl: Chapman & Hall/CRC Press, 2004.

[61] Gilmour, S.G., and Mead, R., "Stopping rules for sequences of factorial designs,", *Applied Statistics*, 44(3), 343–355, 1995.

[62] Giovannitti-Jensen, A., and Myers, R.H., "Graphical assessment of the prediction capability of response surface designs" *Technometrics*, 31, pp. 159–171, 1989.

[63] Goodwin, G.C., and Payne, R.L., *Dynamic System Identification: Experiment Design and Data Analysis*. NY: Academic Press, 1977.

[64] Goos, P., *The Optimal Design of Blocked and Split-Plot Designs*, NY, Springer Verlag, 2004.

[65] Harrington, E.C., "The desirability function", *Industrial Quality Control*, 21, pp. 494–498, 1965.

[66] Hecht-Nielsen, R., *Neurocomputing*, Addison-Wesley, Reading, Mass., 1990.

[67] Heredia-Langner, A., M. Carlyle, D.C. Montgomery, C. Borror, and G. Runger, "Genetic algorithms for the construction of D-optimal designs," *Journal of Quality Technology*, 35, No. 1, pp. 28–46, 2003.

[68] Heredia-Langner, A., D.C. Montgomery, W.M. Carlyle, and C.M. Borror, "Model-robust optimal designs: A genetic algorithm approach", *Journal of Quality Technology*, Vol. 36, No. 3, pp. 263–279, 2004.

[69] Hinkelmann, K., and Kempthorne, O., *Design and Analysis of Experiments*, Volume 2. Wiley, NY, 2005.

[70] Hoerl, R.W., "Ridge Analysis" *Chemical Engineering Symposium Series*, 60, pp. 67–77, 1964.

[71] Hotelling, H., "Experimental determination of the maximum of a function", *Ann. Math. Statist.*, 12, pp. 20–46, 1941.

[72] Joshi, S., Sherali, H.D., and Tew, J.D., "An enhanced Response Surface Methodology (RSM) algorithm using gradient deflection and second order search strategies", *Computers and Ops. Res.*, 25, 7–8, pp. 531–541, 1998.

[73] Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C., "Interactive elicitation of opinion for a normal linear model", *Journal of the American Statistical Association*, 75, pp. 845–854, 1980.

[74] Kass, R.E., and Wasserman, L., "The selection of prior distributions by formal rules", *Journal of the American Statistical Association*, 91, 435, pp. 1343–1370, 1996.

[75] Khuri, A.I., and Conlon, M., "Simultaneous optimization of multiple responses represented by polynomial regression functions", *Technometrics*, 23, pp. 363–375, 1981.

[76] Khuri, A.I., and Cornell, J.A., *Response Surfaces*, 2nd ed., Marcel Dekker, NY, 1996.

[77] Kiefer, J., "Optimum experimental designs", *Journal of the Royal Statistical Society, B (Methodological)*, 21, 2, pp. 272–319, 1959.

[78] Kiefer, J., "Optimal design: Variation in structure and performance under change of criterion", *Biometrika*, 62, 2, pp. 277–288, 1975.

[79] Kiefer, J., and Wolfowtiz, J., "Stochastic estimation of the maximum of a regression function". *Ann. Math. Stat.* 23, 452–466, 1952.

[80] Kiefer, J., and Wolfowitz, J., "Optimum design in regression problems", *Annals of Math. Statist.*, 30, pp. 271–294, 1959.

[81] Kleijnen, J.P.C., Den Hertog, D., Angun, E., "Response Surface Methodology's steepest ascent and step size revisited", *European Journal of Operational Research*, 159, pp. 121–131, 2004.

[82] Kouvelis, P., *Robust Discrete Optimization and Its Applications*, NY: Springer, 2003.

[83] Law, A.M., and Kelton, W.D., *Simulation Modeling and Analysis*, 3rd. ed,, Boston: McGraw Hill, 2000.

[84] Leitsinger, J.D., Myers, R.H., and Lentner, M., "Response Surface methods for Bi-randomization structures", *Journal of Quality Technology*, 28, 4, pp. 381–397, 1996.

[85] Lin, D.K.J., "A new class of supersaturated designs", *Technometrics* 35, pp. 28–31, 1993.

[86] Lophaven, S.N., Nielsen, H.B., Sondergaard, J., "DACE: A Matlab Krigin Toolbox", http://www2.imm.dtu.dk/ hbn/dace/, 2002.

[87] Lucas, J.M., "Optimum composite designs", *Technometrics*, 16, pp. 561–567, 1974.

[88] Lucas, J.M., "Which Response Surface design is best", *Technometrics*, 18, pp. 411–417, 1976.

[89] Lucas, J.M., "How to achieve a robust process using Response Surface Methodology", *Journal of Quality Technology*, 26, pp. 248–260, 1994.

[90] Lucas, J.M., "Letter to the Editor", *Journal of Quality Technology*, 39, 1, pp. 90–91, 2007.

[91] Luenberger, D.G., *Linear and nonlinear programming*, NY: Addison-Wesley, 1984.

[92] Margolin, B.H., "Resolution IV fractional factorial designs", *J. Roy. Statis. Soc. (B)*, 31, pp. 514–523, 1969.

[93] McDaniel, W.R., and Ankenman, B.E., "A Response Surface test bed", *Quality and Reliability Engineering International*, Vol. 16, pp. 363–372, 2000.

[94] McLean, R.A., and Anderson, V.L., "Extreme vertices design of mixture experiments", *Technometrics*, 8, pp. 447–454, 1966.

[95] Mead, R., *The Design of Experiments*, NY: Cambridge University Press, 1988.

[96] Mee, R., "New Box-Behnken Designs", Technical Paper, Dept. of Statistics, University of Tennessee, http://citeseer.ist.psu.edu/625712.html (2000).

[97] Meinhold, R.J., and Singpurwalla N.D., "Understanding the Kalman Filter," *The American Statistician*, 37, 2, 123–127, 1983.

[98] Meyer, R.D., and Box, G., "Finding the Active Factors in Fractionated Screening Experiments", *CQPI Report No. 80*, 1992.

[99] Meyer, R.D., and Box, G., "Finding the active factors in fractionated screening experiments", *Journal of Quality Technology*, 25(2), 94–105, 1993.

[100] Michaels, S.E., "The usefulness of experimental design", *Applied Statistics*, 13, 3, pp. 221–235, 1964.

[101] Miró-Quesada, *Topics in the Optimization of Response Surfaces and Parameter Design Problems*. Ph.D Dissertation, Industrial and Manufacturing Engineering Department, The Pennsylvania State University, 2003.

[102] Miró-Quesada, G., and del Castillo, E., "An enhanced recursive stopping rule for steepest ascent searches in Response Surface Methodology," *Communications in Statistics, Simulation and Computation*, 33, 1, pp. 201–228, 2004.

[103] Miró-Quesada, G., and del Castillo, E., "A search method for the exploration of new regions in robust parameter design," in A. Khuri (ed.) *Advances in Response Surface Methodology*, World Scientific Publishing, 2006.

[104] Miró-Quesada, G., and del Castillo, E., "Two approaches for enhancing the dual response approach to Robust Parameter Design", *Journal of Quality Technology*, 36, 2, pp. 154–168, 2004.

[105] Miró-Quesada, G., and del Castillo, E., "A dual response approach to the multivariate robust parameter design problem", *Technometrics*, 46,2, pp. 176–187, 2004.

[106] Miró-Quesada, G., del Castillo, E., and Peterson, J.J., "A Bayesian approach for multiple Response Surface optimization in the presence of noise variables", *Journal of Applied Statistics*, 31(3), pp. 251–270, 2004.

[107] Mitchell, J., "An algorithm for construction of "D-optimal" experimental designs", *Technometrics*, 16, 2, pp. 203–210, 1974.

[108] Mitchell, J., "Computer construction of "D-optimal" first order designs", *Technometrics*, 16, 2, pp. 211–220, 1974.

[109] Montgomery, D.C., "Experimental design for product and process design and development", *Journal of the Royal Statistical Society (D), The Statistician*, 38, pp. 159–177, 1999.

[110] Montgomery, D.C., *Design and Analysis of Experiments*, 5th ed., New York: Wiley, 2001.

[111] Montgomery, D.C., Peck, E.A., and Vining, G.G., *Introduction to Linear Regression Analysis*, 4th ed., NY: Wiley, 2006.

[112] Myers, R.H., *Response Surface Methodology*, Boston : Allyn and Bacon, 1971.

[113] Myers, R.H., "Response Surface methodology in quality improvement", *Communications in Statistics, Theory and Methods*, 20, 2, pp. 457–476, 1991.

[114] Myers, R.H., and Carter, W.H. Jr., "Response Surface techniques for dual response systems", *Technometrics*, 15, pp. 301–317, 1973.

[115] Myers, R.H., and Khuri, A.I., "A new procedure for steepest ascent," *Communications in Statistics, Theory and Methods*, A8(14), 1359–1376, 1979.

[116] Myers, R.H., Khuri, A.I., and Vining, G., "Response Surface alternatives to the Taguchi robust parameter design approach", *The American Statistician*, 46, 2, 131–139, 1992.

[117] Myers, R.H., and Montgomery, D.C., *Response Surface Methodology*, 2nd ed., New York, Wiley, 2002.

[118] Nair, V.N., (editor), "Taguchi's parameter design: A panel discussion", *Technometrics*, 34, 2, pp. 127–161, 1992.

[119] Nelder, J.A., and Mead, R., "A simplex method for function minimization", *Computer Journal*, 7, p. 308, 1965.

[120] Neyman, J., Review of "Theory of Probability", by H. Jeffreys. *Journal of the American Statistical Association*, 35(211), 558–559, 1940.

[121] O'Reilly, F., "On a criterion for extrapolation in normal regression", *Annals of Statistics*, 3, 1, pp. 219–222, 1975.

[122] Papadimitriou, C.H., *Combinatorial Optimization : Algorithms and Complexity*, NY: Prentice Hall, 1982.

[123] Perez-Salvador, B.R., and O'Reilly, F.J., "Some modifications in Response Surface methodology", Technical Paper, IIMAS, UNAM, Mexico City, 2006.

[124] Peterson, J.J., "A general approach to Ridge Analysis with confidence intervals," *Technometrics*, 35, pp. 204–214, 1993.

[125] Peterson, J.P., "A Bayesian reliability approach to multiple Response Surface optimization", *Journal of Quality Technology*, 36(2), 139–153, 2004.

[126] Peterson, J.J., Cahya, S., and del Castillo, E., "A general approach to confidence regions for optimal factor levels of Response Surfaces", *Biometrics*, 58, pp. 422–431, 2002.

[127] Pignatiello, Jr., J.J., "Strategies for robust multiresponse quality engineering", *IIE Transactions*, 25, pp. 5–15, 1993.

[128] Plackett, R.L. and Burman, J.P., "The design of optimum multifactorial experiments", *Biometrika*, 33, pp. 305–325, 1946.

[129] Press, S.J., *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Robert E. Krieger Publishing, Malabar, FL, 1982.

[130] Press, S.J., *Subjective and Objective Bayesian Statistics*, 2nd ed., New York: John Wiley & Sons, 2003.

[131] Ragajopal, R., and del Castillo, E., "Model-robust process optimization using Bayesian model averaging", *Technometrics*, 47, 2, pp. 152–163, 2005.

[132] Rajagopal, R., and del Castillo, E., "A Bayesian method for robust tolerance control and parameter design," *IIE Transactions*, 38, 8, pp. 685–697, 2007.

[133] Rajagopal, R., del Castillo, E., and Peterson, J.J., "Model and distribution-robust optimization with noise factors,", *Journal of Quality Technology*, 37, 3, pp. 210–222 2005, Corrigenda, 38, 1, pp. 83, 2006.

[134] Rao, C.R., *Linear Statistical Inference and Its Applications,* 2nd ed., NY: Wiley, 1973.

[135] Robbins, H., and Monro, S., "A stochastic approximation method", *Annals of Mathematical Statistics*, 22, pp. 400–407, 1951.

[136] Romano, D., Varetto, M., and Graziam V., "Multiresponse robust design: A general framework based on combined array", *Journal of Quality Technology*, 36, 1, pp. 27–37, 2004.

[137] Roquemore, K.G., "Hybrid designs for quadratic response surfaces", *Technometrics*, 18, pp. 418–423, 1976.

[138] Ross, P.J., *Taguchi techniques for Quality Engineering : Loss Function, Orthogonal Experiments, Parameter and Tolerance Design*, 2nd ed. NY, McGraw-Hill, 1996.

[139] Santner, T.J., Williams, B.J., Notz, W., *The Design and Analysis of Computer Experiments*, NY: Springer Verlag, 2003.

[140] Searle, S.R., *Linear Models*, NY: Wiley, 1971.

[141] Searle, S.R., Casella, G., and McCulloch, C.E., *Variance Components*, NY, Wiley, 1992.

[142] Shoemaker, A.C., Tsui, K.L., and Wu, C.F.J., "Economical experimentation methods for robust design", *Technometrics*, 33, pp. 415–427, 1991.

[143] Bratley, P., Fox, B.L., and Schrage, L.E., *A Guide to Simulation*, 2nd ed., NY: Springer Verlag, 1987.

[144] Spall, J.C., "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, pp. 332–341, 1992.

[145] Spall, J.C., *Introduction to Stochastic Search and Optimization*, John Wiley, NY, 2003.

[146] Stablein, D.L., Carter, W.H., and Wampler, G.L., "Confidence regions for constrained optima in Response Surface experiments," *Biometrics*, 39, pp. 759–763, 1983.

[147] Steuer, R.E., *Multiple Criteria Optimization : Theory, Computation, and Application*, NY: Wiley, 1986.

[148] St. John, R.C., and Draper, N.R., "D-optimality for regression designs: A review", *Technometrics*. 17, pp. 15–23, 1975.

[149] Taguchi, G., *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*, UNIPUB/Kraus International, White Plains, NY, 1987.

[150] Tocher, K.D., "A note on the design problem", *Biometrika*, 39, 1/2, p. 189, 1952.

[151] Trinca, L.A., and Gilmour, S.G., "Multistratum Response Surface designs", *Technometrics*, 43, 1, pp. 25–33, 2001.

[152] Tsui, K.L., "An overview of Taguchi method and newly developed statistical methods for robust design", *IIE Transactions*, 24, 5, pp. 44–57, 1992.

[153] Van Beer, W.C.M., and Kleijnen, J.P.C., "Kriging for interpolation in random simulation", *Journal of the Operational Research Society*, 54, pp. 255–262, 2003.

[154] Vining, G., and Myers, R.H., "A graphical approach for evaluating response surface designs in terms of the mean squared error of prediction", *Technometrics*, 33, 3, pp. 315–326, 1991.

[155] Wei, C.Z., "Multivariate adaptive stochastic approximation," Ann Stat. 15, 1115–1130, 1987.

[156] Wu, C.J.F., "Efficient sequential designs with binary data," JASA, 80, 974–984, 1985.

[157] Wu, C.J.F., and Hamada, M. *Experiments, Planning, Analysis, and Parameter Design Optimization.* Wiley, NY 2000.

[158] Wynn, H.P., "Jack Kiefer's contributions to experimental design", *The Annals of Statistics*, 12, 2, pp. 416–423, 1984.

[159] Yin, X., and Seymour, L., "Standard errors for the multiple roots in quadratic Response Surface Models", *Technometrics*, 47, 3, pp. 260–263, 2005.

[160] Xu, D., and Albin, S.L., "Robust optimization of experimentally derived objective functions", *IIE Transactions*, 35, 9, pp. 793–802, 2003.

[161] Zellner, A., "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions", *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, eds. Goel, P.K., and Zellner, A., Amsterdam, North-Holland, 1986.

[162] Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, New York, Wiley, 1971.

Index

Early Titles in the
INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE

Frederick S. Hillier, Series Editor, *Stanford University*

Saigal/ *A MODERN APPROACH TO LINEAR PROGRAMMING*

Nagurney/ *PROJECTED DYNAMICAL SYSTEMS & VARIATIONAL INEQUALITIES WITH APPLICATIONS*

Padberg & Rijal/ *LOCATION, SCHEDULING, DESIGN AND INTEGER PROGRAMMING*

Vanderbei/ *LINEAR PROGRAMMING*

Jaiswal/ *MILITARY OPERATIONS RESEARCH*

Gal & Greenberg/ *ADVANCES IN SENSITIVITY ANALYSIS & PARAMETRIC PROGRAMMING*

Prabhu/ *FOUNDATIONS OF QUEUEING THEORY*

Fang, Rajasekera & Tsao/ *ENTROPY OPTIMIZATION & MATHEMATICAL PROGRAMMING*

Yu/ *OR IN THE AIRLINE INDUSTRY*

Ho & Tang/ *PRODUCT VARIETY MANAGEMENT*

El-Taha & Stidham/ *SAMPLE-PATH ANALYSIS OF QUEUEING SYSTEMS*

Miettinen/ *NONLINEAR MULTIOBJECTIVE OPTIMIZATION*

Chao & Huntington/ *DESIGNING COMPETITIVE ELECTRICITY MARKETS*

Weglarz/ *PROJECT SCHEDULING: RECENT TRENDS & RESULTS*

Sahin & Polatoglu/ *QUALITY, WARRANTY AND PREVENTIVE MAINTENANCE*

Tavares/ *ADVANCES MODELS FOR PROJECT MANAGEMENT*

Tayur, Ganeshan & Magazine/ *QUANTITATIVE MODELS FOR SUPPLY CHAIN MANAGEMENT*

Weyant, J./ *ENERGY AND ENVIRONMENTAL POLICY MODELING*

Shanthikumar, J.G. & Sumita, U./ *APPLIED PROBABILITY AND STOCHASTIC PROCESSES*

Liu, B. & Esogbue, A.O./ *DECISION CRITERIA AND OPTIMAL INVENTORY PROCESSES*

Gal, T., Stewart, T.J., Hanne, T./ *MULTICRITERIA DECISION MAKING: Advances in MCDM Models, Algorithms, Theory, and Applications*

Fox, B.L./ *STRATEGIES FOR QUASI-MONTE CARLO*

Hall, R.W./ *HANDBOOK OF TRANSPORTATION SCIENCE*

Grassman, W.K./ *COMPUTATIONAL PROBABILITY*

Pomerol, J-C. & Barba-Romero, S./ *MULTICRITERION DECISION IN MANAGEMENT*

Axsater, S./ *INVENTORY CONTROL*

Wolkowicz, H., Saigal, R., & Vandenberghe, L./ *HANDBOOK OF SEMI-DEFINITE PROGRAMMING: Theory, Algorithms, and Applications*

Hobbs, B.F. & Meier, P./ *ENERGY DECISIONS AND THE ENVIRONMENT: A Guide to the Use of Multicriteria Methods*

Dar-El, E./ *HUMAN LEARNING: From Learning Curves to Learning Organizations*

Armstrong, J.S./ *PRINCIPLES OF FORECASTING: A Handbook for Researchers and Practitioners*

Balsamo, S., Personé, V., & Onvural, R./ *ANALYSIS OF QUEUEING NETWORKS WITH BLOCKING*

Bouyssou, D. et al./ *EVALUATION AND DECISION MODELS: A Critical Perspective*

Hanne, T./ *INTELLIGENT STRATEGIES FOR META MULTIPLE CRITERIA DECISION MAKING*

Saaty, T. & Vargas, L./ *MODELS, METHODS, CONCEPTS and APPLICATIONS OF THE ANALYTIC HIERARCHY PROCESS*

Chatterjee, K. & Samuelson, W./ *GAME THEORY AND BUSINESS APPLICATIONS*

Hobbs, B. et al./ *THE NEXT GENERATION OF ELECTRIC POWER UNIT COMMITMENT MODELS*

Vanderbei, R.J./ *LINEAR PROGRAMMING: Foundations and Extensions, 2nd Ed.*

Kimms, A./ *MATHEMATICAL PROGRAMMING AND FINANCIAL OBJECTIVES FOR SCHEDULING PROJECTS*

Baptiste, P., Le Pape, C. & Nuijten, W./ *CONSTRAINT-BASED SCHEDULING*

Feinberg, E. & Shwartz, A./ *HANDBOOK OF MARKOV DECISION PROCESSES*: *Methods and Applications*

Ramík, J. & Vlach, M./ *GENERALIZED CONCAVITY IN FUZZY OPTIMIZATION AND DECISION ANALYSIS*

Song, J. & Yao, D./ *SUPPLY CHAIN STRUCTURES: Coordination, Information and Optimization*

Kozan, E. & Ohuchi, A./ *OPERATIONS RESEARCH/ MANAGEMENT SCIENCE AT WORK*

Bouyssou et al./ *AIDING DECISIONS WITH MULTIPLE CRITERIA: Essays in Honor of Bernard Roy*

Cox, Louis Anthony, Jr./ *RISK ANALYSIS: Foundations, Models and Methods*

Dror, M., L'Ecuyer, P. & Szidarovszky, F./ *MODELING UNCERTAINTY: An Examination of Stochastic Theory, Methods, and Applications*

Dokuchaev, N./ *DYNAMIC PORTFOLIO STRATEGIES: Quantitative Methods and Empirical Rules for Incomplete Information*

Sarker, R., Mohammadian, M. & Yao, X./ *EVOLUTIONARY OPTIMIZATION*

Demeulemeester, R. & Herroelen, W./ *PROJECT SCHEDULING: A Research Handbook*

Gazis, D.C./ *TRAFFIC THEORY*

Zhu/ *QUANTITATIVE MODELS FOR PERFORMANCE EVALUATION AND BENCHMARKING*

Ehrgott & Gandibleux/ *MULTIPLE CRITERIA OPTIMIZATION: State of the Art Annotated Bibliographical Surveys*

Bienstock/ *Potential Function Methods for Approx. Solving Linear Programming Problems*

Matsatsinis & Siskos/ *INTELLIGENT SUPPORT SYSTEMS FOR MARKETING DECISIONS*

Alpern & Gal/ *THE THEORY OF SEARCH GAMES AND RENDEZVOUS*

Hall/ *HANDBOOK OF TRANSPORTATION SCIENCE - 2^{nd} Ed.*

Glover & Kochenberger/ *HANDBOOK OF METAHEURISTICS*

Graves & Ringuest/ *MODELS AND METHODS FOR PROJECT SELECTION: Concepts from Management Science, Finance and Information Technology*

Hassin & Haviv/ *TO QUEUE OR NOT TO QUEUE: Equilibrium Behavior in Queueing Systems*

Gershwin et al/ *ANALYSIS & MODELING OF MANUFACTURING SYSTEMS*

Maros/ *COMPUTATIONAL TECHNIQUES OF THE SIMPLEX METHOD*

Harrison, Lee & Neale/ *THE PRACTICE OF SUPPLY CHAIN MANAGEMENT: Where Theory and Application Converge*

Shanthikumar, Yao & Zijm/ *STOCHASTIC MODELING AND OPTIMIZATION OF MANUFACTURING SYSTEMS AND SUPPLY CHAINS*

Nabrzyski, Schopf & Weglarz/ *GRID RESOURCE MANAGEMENT: State of the Art and Future Trends*

Thissen & Herder/ *CRITICAL INFRASTRUCTURES: State of the Art in Research and Application*

Carlsson, Fedrizzi, & Fullér/ *FUZZY LOGIC IN MANAGEMENT*

Soyer, Mazzuchi & Singpurwalla/ *MATHEMATICAL RELIABILITY: An Expository Perspective*

Chakravarty & Eliashberg/ *MANAGING BUSINESS INTERFACES: Marketing, Engineering, and Manufacturing Perspectives*

Talluri & van Ryzin/ *THE THEORY AND PRACTICE OF REVENUE MANAGEMENT*

Kavadias & Loch/ *PROJECT SELECTION UNDER UNCERTAINTY: Dynamically Allocating Resources to Maximize Value*

Brandeau, Sainfort & Pierskalla/ *OPERATIONS RESEARCH AND HEALTH CARE: A Handbook of Methods and Applications*

Cooper, Seiford & Zhu/ *HANDBOOK OF DATA ENVELOPMENT ANALYSIS: Models and Methods*

Luenberger/ *LINEAR AND NONLINEAR PROGRAMMING, 2^{nd} Ed.*

Early Titles in the
INTERNATIONAL SERIES IN OPERATIONS RESEARCH & MANAGEMENT SCIENCE
(Continued)

Sherbrooke/ *OPTIMAL INVENTORY MODELING OF SYSTEMS: Multi-Echelon Techniques, Second Edition*

Chu, Leung, Hui & Cheung/ *4th PARTY CYBER LOGISTICS FOR AIR CARGO*

Simchi-Levi, Wu & Shen/ *HANDBOOK OF QUANTITATIVE SUPPLY CHAIN ANALYSIS: Modeling in the E-Business Era*

**** A list of the more recent publications in the series is at the front of the book ****