

# Springer Series in Statistics

*Advisors:*

P. Bickel, P. Diggle, S. Fienberg, U. Gather,

I. Olkin, S. Zeger

# Springer Series in Statistics

---

- Alho/Spencer*: Statistical Demography and Forecasting  
*Andersen/Borgan/Gill/Keiding*: Statistical Models Based on Counting Processes  
*Atkinson/Riani*: Robust Diagnostic Regression Analysis  
*Atkinson/Riani/Cerilo*: Exploring Multivariate Data with the Forward Search  
*Berger*: Statistical Decision Theory and Bayesian Analysis, 2<sup>nd</sup> edition  
*Borg/Groenen*: Modern Multidimensional Scaling: Theory and Applications, 2<sup>nd</sup> edition  
*Brockwell/Davis*: Time Series: Theory and Methods, 2<sup>nd</sup> edition  
*Bucklew*: Introduction to Rare Event Simulation  
*Cappé/Moulines/Rydén*: Inference in Hidden Markov Models  
*Chan/Tong*: *Chaos: A Statistical Perspective*  
*Chen/Shao/Ibrahim*: Monte Carlo Methods in Bayesian Computation  
*Coles*: An Introduction to Statistical Modeling of Extreme Values  
*Devroye/Lugosi*: Combinatorial Methods in Density Estimation  
*Diggel/Ribeiro*: Model-based Geostatistics  
*Dudoit/Van der Laan*: Multiple Testing Procedures with Applications to Genomics  
*Efromovich*: Nonparametric Curve Estimation: Methods, Theory, and Applications  
*Eggermont/LaRiccia*: Maximum Penalized Likelihood Estimation, Volume I: Density Estimation  
*Fahrmeir/Tutz*: Multivariate Statistical Modeling Based on Generalized Linear Models, 2<sup>nd</sup> edition  
*Fan/Yao*: *Nonlinear Time Series: Nonparametric and Parametric Methods*  
*Ferraty/Vieu*: Nonparametric Functional Data Analysis: Theory and Practice  
*Ferreira/Lee*: Multiscale Modeling: A Bayesian Perspective  
*Fienberg/Hoaglin*: Selected Papers of Frederick Mosteller  
*Frühwirth-Schnatter*: Finite Mixture and Markov Switching Models  
*Ghosh/Ramamoorthi*: Bayesian Nonparametrics  
*Glaz/Naus/Wallenstein*: Scan Statistics  
*Good*: Permutation Tests: Parametric and Bootstrap Tests of Hypotheses, 3rd edition  
*Gouriéroux*: ARCH Models and Financial Applications  
*Gu*: Smoothing Spline ANOVA Models  
*Gyöfi/Kohler/Krzyżak/Walk*: A Distribution-Free Theory of Nonparametric Regression  
*Haberman*: Advanced Statistics, Volume I: Description of Populations  
*Hall*: The Bootstrap and Edgeworth Expansion  
*Härdle*: Smoothing Techniques: With Implementation in S  
*Harrell*: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis  
*Hart*: Nonparametric Smoothing and Lack-of-Fit Tests  
*Hastie/Tibshirani/Friedman*: The Elements of Statistical Learning: Data Mining, Inference, and Prediction  
*Hedayat/Sloanel/Stufken*: Orthogonal Arrays: Theory and Applications  
*Heyde*: Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation  
*Huet/Bouvier/Poursat/Jolivet*: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples, 2<sup>nd</sup> edition  
*Ibrahim/Chen/Sinha*: Bayesian Survival Analysis  
*Jiang*: Linear and Generalized Linear Mixed Models and Their Applications  
*Jolliffe*: Principal Component Analysis, 2<sup>nd</sup> edition  
*Knottnerus*: Sample Survey Theory: Some Pythagorean Perspectives

(continued after index)

Peter X.-K. Song

# Correlated Data Analysis: Modeling, Analytics, and Applications



Springer

Peter X.-K. Song  
Department of Statistics and Actuarial Science  
University of Waterloo  
200 University Avenue West  
Waterloo, Ontario, Canada N2L 3G1  
song@uwaterloo.ca

Library of Congress Control Number: 2007929730

ISBN 978-0-387-71392-2

e-ISBN 978-0-387-71393-9

Printed on acid-free paper.

©2007 Springer Science+Business Media, LLC

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

To Ru and Angela

---

## Preface

This book, like many other books, was delivered under tremendous inspiration and encouragement from my teachers, research collaborators, and students. My interest in longitudinal data analysis began with a short course taught jointly by K.Y. Liang and S.L. Zeger at the Statistical Society of Canada Conference in Acadia University, Nova Scotia, in the spring of 1993. At that time, I was a first-year PhD student in the Department of Statistics at the University of British Columbia, and was eagerly seeking potential topics for my PhD dissertation. It was my curiosity (driven largely by my terrible confusion) with the generalized estimating equations (GEEs) introduced in the short course that attracted me to the field of correlated data analysis. I hope that my experience in learning about it has enabled me to make this book an enjoyable intellectual journey for new researchers entering the field. Thus, the book aims at graduate students and methodology researchers in statistics or biostatistics who are interested in learning the theory and methods of correlated data analysis.

I have attempted to give a systematic account of regression models and their applications to the modeling and analysis of correlated data. Longitudinal data, as an important type of correlated data, has been used as a main venue for motivation, methodological development, and illustration throughout the book. Given the many applied books on longitudinal data analysis already available, this book is inclined more towards technical details regarding the underlying theory and methodology used in software-based applications. I hope the book will serve as a useful reference for those who want theoretical explanations to puzzles arising from data analyses or deeper understanding of underlying theory related to analyses. This book has evolved from lecture notes on longitudinal data analysis, and may be considered suitable as a textbook for a graduate course on correlated data analysis.

This book emphasizes some recent developments in correlated data analysis.

First, it takes the perspective of Jørgensen's theory of dispersion models for the discussion of generalized linear models (GLMs) in Chapter 2. It

is known that the class of generalized linear models plays a central role in the regression analysis of nonnormal data. In the context of correlated data analysis, these models constitute marginal components in a joint model formulation. One benefit from such a treatment is that it enables this book to cover a broader range of data types than the traditional GLMs. Two types that are of particular interest and discussed in detail in the book are compositional (or continuous proportional) data and directional (or circular) data.

Second, it gives a systematic treatment for the theory of inference functions (or estimating functions) in Chapter 3. The popular GEE methods presented in Chapter 5 are then easily introduced and studied as a special class of inference functions. Building upon Chapter 3, some alternative estimating function methods can be readily discussed. Recent work on quadratic inference functions (QIF) is an example that benefits from Chapter 3.

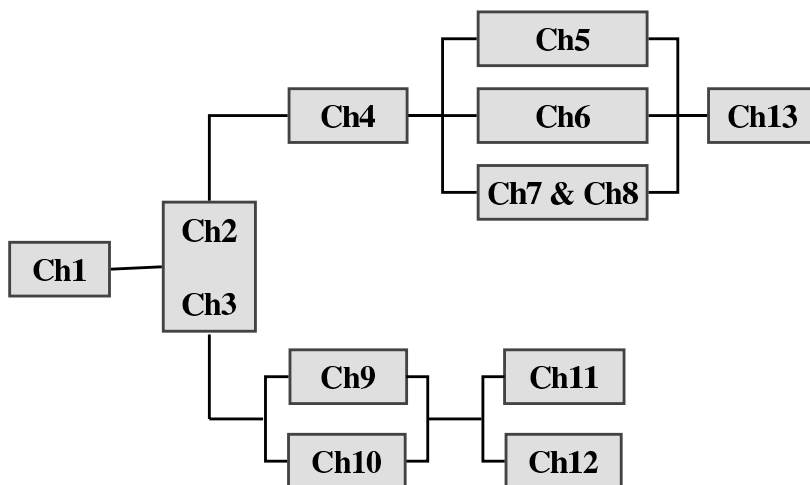
Third, it presents a joint modeling approach to regression analysis of correlated data via the technique of parametric copulas. Copulas are becoming increasingly popular in the analysis of correlated data, and Chapter 6 focuses on Gaussian copulas, for which both theory and numerical examples are illustrated.

Fourth, it deals with state space models for longitudinal data from long time series. In contrast to longitudinal data from short time series, modeling stochastic patterns or transitional behaviors becomes a primary task. In such a setting, asymptotics may be established by letting the length of the time series tend to  $\infty$ , as opposed to letting the number of subjects tend to  $\infty$ , as in the case of data consisting of many short time series. Chapters 10, 11, and 12 are devoted to this topic.

Fifth, this book covers two kinds of statistical inferences in generalized linear mixed effects models (GLMMs): maximum likelihood inference in Chapter 7 and Bayesian inference based on Markov Chain Monte Carlo (MCMC) in Chapter 8. In Chapter 8, the analysis of multi-level data is also discussed in the framework of hierarchical models. Inference can be dealt with easily by the MCMC method, as an extension from the GLMMs with little extra technical difficulty.

The book contains some other topics that are highly relevant to the analysis of correlated data. For example, Chapter 13 concerns missing data problems arising particularly from longitudinal data.

The presentation of some material in the book is a little technical in order to achieve rigor of exposition. Readers' backgrounds should include mathematical statistics, generalized linear models, and some knowledge of statistical computing, such as represented R and SAS software. The following chart displays the relationship among the thirteen chapters, and readers can follow a particular path to reach a topic of interest.



A webpage has been created to provide some supplementary material for the book. The URL address is

<http://www.stats.uwaterloo.ca/~song/BOOKLDA.html>

All data sets used in the book are available. A SAS Macro QIF is available for a secured download; that is, an interested user needs to submit an online request for permission in order to download this software package. In addition, some figures that are printed in reduced size in the book are supplied in their full sizes. Exercise problems for some of the thirteen chapters are posted, which may be useful when the book is used as a text for a course.

I would like to acknowledge my debt to many people who have helped me to prepare the book. I was fortunate to begin my research in this field under the supervision of Bent Jørgensen, who taught me his beautiful theory of dispersion models. At UBC, I learned the theory of copulas from Harry Joe. This book has benefited from some of the PhD theses that I supervised in the past ten years or so, including Zhenguo (Winston) Qiu, Dingan Feng, Baifang Xing, and Peng Zhang, as well as from a few data analysis projects that graduate students did in my longitudinal data analysis course; thanks go to Eric Bingshu Chen, Wenyu Jiang, David Tolusso, and Wanhua Su. Many graduate students in my course pointed out errors in an early draft of the book. Qian Zhou helped me to draw some figures in the book, and Zichang Jiang worked with me to develop SAS MACRO QIF, which is a software package to fit marginal models for correlated data.

I am very grateful to my research collaborators for their constant inspiration and valuable discussions on almost every topic presented in the book. My great appreciation goes to Annie Qu, Jack Kalbfleisch, Ming Tan, Claudia Czado, Søren Lundbye-Christensen, Jianguo (Tony) Sun, and Mingyao Li. I would also like to express my sincere gratitude to people who generously provided and allowed me to analyze their datasets in the book, including John



Petkau and Angela D'Elia. Zhenguo Qiu, Grace Yi, and Jerry Lawless provided with me their valuable comments on drafts of the book. My research in the field of correlated data analysis has been constantly supported by grants from the Natural Sciences and Engineering Research Council of Canada. I thank John Kimmel and Frank Ganz from Springer for their patience and editorial assistance.

I take full responsibility for all errors and omissions in the book. Finally, I would like to say that given the vast amount of published material in the field of correlated data analysis, the criterion that I adopted for the selection of topics for the book was really my own familiarity. Because of this and space limitations, some worthwhile topics have no doubt been excluded. Research in this field remains very active with many new developments. I would be grateful to readers for their critical comments and suggestions for improvement, as well as corrections.

Waterloo, Ontario, Canada

*P.X.-K. Song*  
December 2006

---

# Contents

<b>Preface</b> .....	vii
<b>1 Introduction and Examples</b> .....	1
1.1 Correlated Data .....	1
1.2 Longitudinal Data Analysis .....	2
1.3 Data Examples .....	6
1.3.1 Indonesian Children's Health Study .....	6
1.3.2 Epileptic Seizures Data .....	7
1.3.3 Retinal Surgery Data .....	9
1.3.4 Orientation of Sandhoppers .....	10
1.3.5 Schizophrenia Clinical Trial .....	11
1.3.6 Multiple Sclerosis Trial .....	13
1.3.7 Tretinoin Emollient Cream Trial .....	13
1.3.8 Polio Incidences in USA .....	14
1.3.9 Tokyo Rainfall Data .....	15
1.3.10 Prince George Air Pollution Study .....	16
1.4 Remarks .....	19
1.5 Outline of Subsequent Chapters .....	20
<b>2 Dispersion Models</b> .....	23
2.1 Introduction .....	23
2.2 Dispersion Models .....	25
2.2.1 Definitions .....	26
2.2.2 Properties .....	28
2.3 Exponential Dispersion Models .....	30
2.4 Residuals .....	35
2.5 Tweedie Class .....	36
2.6 Maximum Likelihood Estimation .....	37
2.6.1 General Theory .....	38
2.6.2 MLE in the ED Models .....	41
2.6.3 MLE in the Simplex GLM .....	42

2.6.4	MLE in the von Mises GLM .....	49
<b>3</b>	<b>Inference Functions</b> .....	<b>55</b>
3.1	Introduction .....	55
3.2	Quasi-Likelihood Inference in GLMs .....	56
3.3	Preliminaries.....	58
3.4	Optimal Inference Functions .....	61
3.5	Multi-Dimensional Inference Functions .....	65
3.6	Generalized Method of Moments .....	68
<b>4</b>	<b>Modeling Correlated Data</b> .....	<b>73</b>
4.1	Introduction .....	73
4.2	Quasi-Likelihood Approach .....	76
4.3	Conditional Modeling Approaches .....	80
4.3.1	Latent Variable Based Approach .....	80
4.3.2	Transitional Model Based Approach .....	82
4.4	Joint Modeling Approach .....	84
<b>5</b>	<b>Marginal Generalized Linear Models</b> .....	<b>87</b>
5.1	Model Formulation .....	88
5.2	GEE: Generalized Estimating Equations.....	89
5.2.1	General Theory .....	90
5.2.2	Some Special Cases .....	93
5.2.3	Wald Test for Nested Models .....	95
5.3	GEE2 .....	95
5.3.1	Constant Dispersion Parameter .....	96
5.3.2	Varying Dispersion Parameter .....	100
5.4	Residual Analysis .....	101
5.4.1	Checking Distributional Assumption .....	102
5.4.2	Checking Constant Dispersion Assumption .....	102
5.4.3	Checking Link Functions .....	102
5.4.4	Checking Working Correlation .....	102
5.5	Quadratic Inference Functions.....	103
5.6	Implementation and Softwares .....	106
5.6.1	Newton-Scoring Algorithm .....	106
5.6.2	SAS PROC GENMOD .....	107
5.6.3	SAS MACRO QIF .....	108
5.7	Examples.....	109
5.7.1	Longitudinal Binary Data .....	110
5.7.2	Longitudinal Count Data .....	112
5.7.3	Longitudinal Proportional Data .....	116

<b>6</b>	<b>Vector Generalized Linear Models</b>	121
6.1	Introduction	121
6.2	Log-Linear Model for Correlated Binary Data	122
6.3	Multivariate ED Family Distributions	125
6.3.1	Copulas	126
6.3.2	Construction	127
6.3.3	Interpretation of Association Parameter	129
6.4	Simultaneous Maximum Likelihood Inference	136
6.4.1	General Theory	136
6.4.2	VGLMs for Correlated Continuous Outcomes	137
6.4.3	VGLMs for Correlated Discrete Outcomes	138
6.4.4	Scores for Association Parameters	139
6.5	Algorithms	141
6.5.1	Algorithm I: Maximization by Parts	142
6.5.2	Algorithm II: Gauss-Newton Type	146
6.6	An Illustration: VGLMs for Trivariate Discrete Data	146
6.6.1	Trivariate VGLMs	147
6.6.2	Comparison of Asymptotic Efficiency	148
6.7	Data Examples	150
6.7.1	Analysis of Two-Period Cross-Over Trial Data	150
6.7.2	Analysis of Hospital Visit Data	152
6.7.3	Analysis of Burn Injury Data	153
<b>7</b>	<b>Mixed-Effects Models: Likelihood-Based Inference</b>	157
7.1	Introduction	157
7.2	Model Specification	161
7.3	Estimation	165
7.4	MLE Based on Numerical Integration	167
7.5	Simulated MLE	174
7.6	Conditional Likelihood Estimation	176
7.7	MLE Based on EM Algorithm	178
7.8	Approximate Inference: PQL and REML	182
7.9	SAS Software	192
7.9.1	PROC MIXED	192
7.9.2	PROC NL MIXED	193
7.9.3	PROC GLIMMIX	194
<b>8</b>	<b>Mixed-Effects Models: Bayesian Inference</b>	195
8.1	Bayesian Inference Using MCMC Algorithm	195
8.1.1	Gibbs Sampling: A Practical View	195
8.1.2	Diagnostics	198
8.1.3	Enhancing Burn-in	201
8.1.4	Model Selection	202
8.2	An Illustration: Multiple Sclerosis Trial Data	203
8.3	Multi-Level Correlated Data	206

8.4	WinBUGS Software	212
8.4.1	WinBUGS Code in Multiple Sclerosis Trial Data Analysis	213
8.4.2	WinBUGS Code for the TEC Drug Analysis	214
<b>9</b>	<b>Linear Predictors</b>	<b>217</b>
9.1	General Results	217
9.2	Estimation of Random Effects in GLMMs	221
9.2.1	Estimation in LMMs	221
9.2.2	Estimation in GLMMs	221
9.3	Kalman Filter and Smoother	222
9.3.1	General Forms	222
<b>10</b>	<b>Generalized State Space Models</b>	<b>227</b>
10.1	Introduction	227
10.2	Linear State Space Models	231
10.3	Shift-Mean Model	232
10.4	Monte Carlo Maximum Likelihood Estimation	235
<b>11</b>	<b>Generalized State Space Models for Longitudinal Binomial Data</b>	<b>239</b>
11.1	Introduction	239
11.2	Monte Carlo Kalman Filter and Smoother	240
11.3	Bayesian Inference Based on MCMC	246
<b>12</b>	<b>Generalized State Space Models for Longitudinal Count Data</b>	<b>261</b>
12.1	Introduction	261
12.2	Generalized Estimating Equation	264
12.3	Monte Carlo EM Algorithm	265
12.4	KEE in Stationary State Processes	267
12.4.1	Setup	267
12.4.2	Kalman Filter and Smoother	269
12.4.3	Godambe Information Matrix	271
12.4.4	Analysis of Polio Incidences Data	272
12.5	KEE in Non-Stationary State Processes	275
12.5.1	Model Formulation	275
12.5.2	Kalman Filter and Smoother	278
12.5.3	Parameter Estimation	280
12.5.4	Model Diagnosis	281
12.5.5	Analysis of Prince George Data	283

**13 Missing Data in Longitudinal Studies** . . . . . 291

13.1 Introduction . . . . . 291

13.2 Missing Data Patterns . . . . . 293

    13.2.1 Patterns of Missingness . . . . . 293

    13.2.2 Types of Missingness and Effects . . . . . 297

13.3 Diagnosis of Missing Data Types . . . . . 300

    13.3.1 Graphic Approach . . . . . 301

    13.3.2 Testing for MCAR . . . . . 302

13.4 Handling MAR Mechanism . . . . . 306

    13.4.1 Simple Solutions and Limitations . . . . . 307

    13.4.2 Multiple Imputation . . . . . 307

    13.4.3 EM Algorithm . . . . . 311

    13.4.4 Inverse Probability Weighting . . . . . 317

13.5 Handling NMAR Mechanism . . . . . 320

    13.5.1 Parametric Modeling . . . . . 320

    13.5.2 A Semiparametric Pattern Mixture Model . . . . . 322

**References** . . . . . 329

**Index** . . . . . 343