Appendix A
Data Sets

All data sets are included in the R library SMSdata that may be downloaded via the quantlet download center: www.quantlet.org. All data sets are available also on the Springer webpage.

A.1 Athletic Records Data

This data set provides data on athletic records in 100, 200, 400, 800, 1,500, 5,000, 10,000 m, and Marathon for 55 countries.

A.2 Bank Notes Data

Six variables were measured on 100 genuine and 100 counterfeit old Swiss 1000-franc bank notes. The data stem from Flury and Riedwyl (1988). The columns correspond to the following 6 variables.

- $X_1$: length of the bank note
- $X_2$: height of the bank note, measured on the left
- $X_3$: height of the bank note, measured on the right
- $X_4$: distance of the inner frame to the lower border
- $X_5$: distance of the inner frame to the upper border
- $X_6$: length of the diagonal

Observations 1–100 are the genuine bank notes and the other 100 observations are the counterfeit bank notes.
A.3  Bankruptcy Data

The data are the profitability, leverage, and bankruptcy indicators for 84 companies. The data set contains information on 42 of the largest companies that filed for protection against creditors under Chap. 11 of the U.S. Bankruptcy Code in 2001–2002 after the stock market crash of 2000. The bankrupt companies were matched with 42 surviving companies with the closest capitalizations and the same US industry classification codes available through the Division of Corporate Finance of the Securities and Exchange Commission (SEC 2004). The information for each company was collected from the annual reports for 1998–1999 (SEC 2004), i.e., 3 years prior to the defaults of the bankrupt companies. The following data set contains profitability and leverage ratios calculated, respectively, as the ratio of net income (NI) and total assets (TA) and the ratio of total liabilities (TL) and total assets (TA).

A.4  Car Data

The car data set (Chambers et al. 1983) consists of 13 variables measured for 74 car types. The abbreviations in the data set are as follows:

- $X_1$: P price
- $X_2$: M mileage (in miles per gallon)
- $X_3$: R78 repair record 1978 (rated on a 5-point scale: 5 best, 1 worst)
- $X_4$: R77 repair record 1977 (scale as before)
- $X_5$: H headroom (in inches)
- $X_6$: R rear seat clearance (in inches)
- $X_7$: Tr trunk space (in cubic feet)
- $X_8$: W weight (in pound)
- $X_9$: L length (in inches)
- $X_{10}$: T turning diameter (clearance required to make a U-turn, in feet)
- $X_{11}$: D displacement (in cubic inches)
- $X_{12}$: G gear ratio for high gear
- $X_{13}$: C company headquarters (1 United States, 2 Japan, 3 Europe)

A.5  Car Marks

The data are averaged marks for 24 car types from a sample of 40 persons. The marks range from 1 (very good) to 6 (very bad) like German school marks. The variables are:

- $X_1$: A economy
- $X_2$: B service
A.6 Classic Blue Pullover Data

This is a data set consisting of 10 measurements of 4 variables. A textile shop manager is studying the sales of “classic blue” pullovers over 10 periods. He uses three different marketing methods and hopes to understand his sales as a fit of these variables using statistics. The variables measured are

- $X_1$: number of sold pullovers
- $X_2$: price (in EUR)
- $X_3$: advertisement costs in local newspapers (in EUR)
- $X_4$: presence of a sales assistant (in hours per period)

A.7 Fertilizer Data

The yields of wheat have been measured in 30 parcels, which have been randomly attributed to 3 lots prepared by one of 3 different fertilizers A, B, and C.

- $X_1$: fertilizer A
- $X_2$: fertilizer B
- $X_3$: fertilizer C

A.8 French Baccalauréat Frequencies

The data consist of observations of 202,100 French baccalauréats in 1976 and give the frequencies for different sets of modalities classified into regions. For a reference, see Bouroche and Saporta (1980). The variables (modalities) are:

- $X_1$: A philosophy letters
- $X_2$: B economics and social sciences
- $X_3$: C mathematics and physics
- $X_4$: D mathematics and natural sciences
- $X_5$: E mathematics and techniques
A.9 French Food Data

The data set consists of the average expenditures on food (bread, vegetables, fruit, meat, poultry, milk, and wine) for several different types of families in France (manual workers = MA, employees = EM, managers = CA) with different numbers of children (2, 3, 4, or 5 family members). The data are taken from Lebart et al. (1982).

A.10 Geopol Data

This data set contains a comparison of 41 countries according to 10 different political and economic parameters:

- $X_1$: popu   population
- $X_2$: giph   gross internal product per habitant
- $X_3$: ripo   rate of increase of the population
- $X_4$: rupo   rate of urban population
- $X_5$: rlpo   rate of illiteracy in the population
- $X_6$: rspo   rate of students in the population
- $X_7$: eltp   expected lifetime of people
- $X_8$: rnmr   rate of nutritional needs realized
- $X_9$: nunh   number of newspapers and magazines per 1,000 habitants
- $X_{10}$: nuth  number of television per 1,000 habitants

A.11 German Annual Population Data

The data set shows yearly average population and unemployment rates for the old federal states in Germany (given in 1,000 inhabitants).

A.12 Journals Data

This is a data set that was created from a survey completed in the 1980’s in Belgium questioning people’s reading habits. They were asked where they live (10 regions
comprising 7 provinces and 3 regions around Brussels) and what kind of newspaper they read on a regular basis. The 15 possible answers belong to 3 classes: Flemish newspapers (first letter v), French newspapers (first letter f) and both languages (first letter b).y

\[ X_1: \quad \text{WaBr} \quad \text{Walloon Brabant} \]

\[ X_2: \quad \text{Brar} \quad \text{Brussels area} \]

\[ X_3: \quad \text{Antw} \quad \text{Antwerp} \]

\[ X_4: \quad \text{FlBr} \quad \text{Flemish Brabant} \]

\[ X_5: \quad \text{OcFl} \quad \text{Occidental Flanders} \]

\[ X_6: \quad \text{OrFl} \quad \text{Oriental Flanders} \]

\[ X_7: \quad \text{Hain} \quad \text{Hainaut} \]

\[ X_8: \quad \text{Lièg} \quad \text{Liège} \]

\[ X_9: \quad \text{Limb} \quad \text{Limburg} \]

\[ X_{10}: \quad \text{Luxe} \quad \text{Luxembourg} \]

A.13 NYSE Returns Data

This data set consists of returns of seven stocks traded on the New York Stock Exchange (Berndt 1990). The monthly returns of IBM, PanAm, Delta Airlines, Consolidated Edison, Gerber, Texaco, and Digital Equipment Company are stated from January 1978 to December 1987.

A.14 Plasma Data

In Olkin and Veath (1980), the evolution of citrate concentration in the plasma is observed at 3 different times of day for two groups of patients. Each group follows a different diet.

\[ X_1: \quad 8\ AM \]

\[ X_2: \quad 11\ AM \]

\[ X_3: \quad 3\ PM \]

A.15 Time Budget Data

In Volle (1985), we can find data on 28 individuals identified according to gender, country where they live, professional activity, and matrimonial status, which indicates the amount of time each person spent on 10 categories of activities over 100 days (100-24 h = 2,400 h total in each row) in 1976.
$X_1$: prof : professional activity
$X_2$: tran : transportation linked to professional activity
$X_3$: hous : household occupation
$X_4$: kids : occupation linked to children
$X_5$: shop : shopping
$X_6$: pers : time spent for personal care
$X_7$: eat : eating
$X_8$: slee : sleeping
$X_9$: tele : watching television
$X_{10}$: leis : other leisure activities
maus: active men in the United States
waus: active women in the United States
wnus: nonactive women in the United States
mmus: married men in United States
wmus: married women in United States
msus: single men in United States
wsus: single women in United States
mawe: active men from Western countries
wawe: active women from Western countries
wnwe: nonactive women from Western countries
mmwe: married men from Western countries
wmwe: married women from Western countries
mswe: single men from Western countries
wswe: single women from Western countries
mayo: active men from Yugoslavia
wayo: active women from Yugoslavia
wnyo: nonactive women from Yugoslavia
mmyo: married men from Yugoslavia
wmyo: married women from Yugoslavia
msyo: single men from Yugoslavia
wsyo: single women from Yugoslavia
maea: active men from Eastern countries
waea: active women from Eastern countries
wnea: nonactive women from Eastern countries
mmea: married men from Eastern countries
wmea: married women from Eastern countries
msea: single men from Eastern countries
wsea: single women from Eastern countries
wsea: single women from Eastern countries


A.16 Unemployment Data

This data set provides unemployment rates in all federal states of Germany in September 1999.

A.17 U.S. Companies Data

The data set consists of measurements for 79 U.S. companies. The abbreviations are as follows:

- $X_1$: A assets (USD)
- $X_2$: S sales (USD)
- $X_3$: MV market value (USD)
- $X_4$: P profits (USD)
- $X_5$: CF cash flow (USD)
- $X_6$: E employees

A.18 U.S. Crime Data

This is a data set consisting of 50 measurements of 7 variables. It states for 1 year (1985) the reported number of crimes in the 50 states of the United States classified according to 7 categories ($X_3$–$X_9$):

- $X_1$: land area (land)
- $X_2$: population 1985 (popu 1985)
- $X_3$: murder (murd)
- $X_4$: rape
- $X_5$: robbery (robb)
- $X_6$: assault (assa)
- $X_7$: burglary (burg)
- $X_8$: larcery (larc)
- $X_9$: auto theft (auto)
- $X_{10}$: U.S. states region number (reg)
- $X_{11}$: U.S. states division number (div)
A.19  U.S. Health Data

This is a data set consisting of 50 measurements of 13 variables. It states for 1 year (1985) the reported number of deaths in the 50 states of the U.S. classified according to 7 categories:

\[
\begin{align*}
X_1 &: \text{land area (land)} \\
X_2 &: \text{population 1985 (popu)} \\
X_3 &: \text{accident (acc)} \\
X_4 &: \text{cardiovascular (card)} \\
X_5 &: \text{cancer (canc)} \\
X_6 &: \text{pulmonary (pul)} \\
X_7 &: \text{pneumonia flu (pneu)} \\
X_8 &: \text{diabetes (diab)} \\
X_9 &: \text{liver (liv)} \\
X_{10} &: \text{doctors (doc)} \\
X_{11} &: \text{hospitals (hosp)} \\
X_{12} &: \text{U.S. states region number (reg)} \\
X_{13} &: \text{U.S. states division number (div)}
\end{align*}
\]

A.20  Vocabulary Data

This example of the evolution of the vocabulary of children can be found in Bock (1975). Data are drawn from test results on file in the Records Office of the Laboratory School of the University of Chicago. They consist of scores, obtained from a cohort of pupils from the 8th through 11th grade levels, on alternative forms of the vocabulary section of the Cooperative Reading Test. It provides scaled scores for the sample of 64 subjects (the origin and units are fixed arbitrarily).
Morrison (1990) compares the results of 4 subtests of the Wechsler Adult Intelligence Scale (WAIS) for 2 categories of people. In group 1 are $n_1 = 37$ people who do not present a senile factor; in group 2 are those ($n_2 = 12$) presenting a senile factor.

**WAIS subtests:**
- $X_1$: information
- $X_2$: similarities
- $X_3$: arithmetic
- $X_4$: picture completion
References


References


References


Index

Actual error rate (AER), 258
Adjoint matrix, xix, 22
Agglomerative algorithm, 225
Akaike's information criterion (AIC), 155, 158
Analysis of variance (ANOVA), 123, 126–128, 144, 301, 302
balanced, 145
unbalanced, 147
ANCOVA, 144
Andrews' curves, 3, 15
ANOVA. See Analysis of variance (ANOVA)
aplpack, viii
Apparent error rate (APER), 251, 258
Asymptotic normality, xix
Average linkage, 226
Bagplot, 320
Balanced dice, 105
Bandwidth, 7
Bartlett correction, 214
Bayes discrimination rule, 246, 250, 252
Beta factor, 311
Between-group sum of squares, 247, 252
Between variance, 246
Bias, xix
Binomial distribution, 257
Binomial inverse theorem, 25
Boxplot, 3, 4, 6, 33, 320

c, viii, 263
Canonical correlation analysis (CCA)
analysis, 281, 283, 287
coefficient, 281, 282
variables, 281, 282
vectors, 281, 282
Canonical hyperplane, 323
Capital asset pricing model (CAPM), 309, 310
car, viii, 161
CART. See Classification and regression tree (CART)
Cauchy distribution, 50
CCA. See Canonical correlation analysis (CCA)
cdf. See Cumulative distribution function (cdf)
Centering matrix, xvii, 36, 74
Central hole index, 321
Central limit theorem (CLT), xviii, 44, 50
Central mass index, 321
Centroid distance, 226
Centroid linkage, 226
Characteristic function, xvi
Characteristic polynomial, xx
Chernoff faces. See Flury–Chernoff faces
χ² distance, 239
χ² distribution, xviii
quantile, xviii
χ² statistic, 259, 260, 266
Classification and regression trees (CART), 319, 322
Gini criterion, 323
least squares criterion, 323
twoing criterion, 323
CLT. See Central limit theorem (CLT)
Cluster analysis, 225
average linkage, 226
centroid linkage, 226
complete linkage, 226, 233
median linkage, 226
Index

single linkage, 226, 230, 233
Ward method, 226, 227, 229, 230, 235
Coefficient of determination, 32, 37, 38, 144
Cofactor, xx
Column coordinates, 263
Column factor, 263, 264
contributions to variance, 264
Column frequency, 259
Common factors, 205
Communality, 206
Complete linkage, 226, 233
Computational statistics, 319
Conditional distribution, xx, 45, 56, 59, 69, 72, 73, 75, 76, 78–80
Conditional expectation, xvi, 46, 47, 53, 55, 58, 69, 79, 81, 85, 86
Conditional moments, xx, 45
Conditional pdf, 58, 62
Conditional probability, 59
Conditional variance, xvi, 47, 53, 58, 81
Confidence interval, 143
Confidence region, 114
Conjoint measurement analysis, 301
nonmetric solution, 302
Consumer preferences, 301
Contingency table, xx, 239, 259, 265, 276–278
Contour ellipse, 66
Contrast matrix, 126, 132
Convergence
in distribution, xviii
in probability, xviii
Convex hull peeling, xviii, 320
Correlation, xvi, 29, 32, 34, 36, 81, 83, 281
empirical, xvii
multiple, 85
partial, 84
Correlation matrix, 29, 36, 180
empirical, xvii
corresp, 263
Correspondence analysis, 259, 260, 276–278
explained variance, 264
Cost of misclassification, 255
Covariance, xvi, 28, 29, 35, 36, 42, 55, 88, 170
empirical, xvii
Covariance matrix, xvi, 37, 42, 43, 55, 72, 183
diagonal, 98
empirical, xvii
partitioned, 71, 88
Covariance test statistic, 163
Cramer-Rao lower bound, 91, 95, 96, 98
Cramer-Rao theorem, 90
Credit scoring, 245
Critical value, xx, 104
Cross-validation, 161
Cumulants, xvi
Cumulative distribution function cdf, xvi, xx, xxiii, 43
empirical, xxi
joint, xvi
marginal, xvi
Data cloud, 167, 168
Data depth, 319
Data matrix, xvii
standardized, 36
Dendrogram, 226
Density. See Probability density function (pdf)
depth, viii
Depth function, 320
Derivative, xx
Descriptive techniques, 3
Design matrix, 141, 301, 304
Determinant, xviii, xxi, 21, 22, 24
Deviance, 145
null, 155
residual, 145, 155
Diagonal, xviii
Discrimination
analysis, 245
Fisher’s (see Fisher’s LDA)
ML (see ML discrimination)
rule, 245
Disparity, 290
Dissimilarity, 290
Distance matrix, 225, 289
Distribution, xv
χ², xviii
conditional, xx, 45
exponential, 69
F-, xviii
Gaussian, xxiii
Hotelling, xviii, 72, 108, 112, 126
marginal, xxii, 46
multinormal, xxiii
normal, xviii, xxiii
t-, xviii
Wishart, 72, 133
Distribution function. See also Cumulative
distribution function (cdf)
empirical, xxi
dr, viii, 322
Draftman plot, 3, 15
Duality, 168
<table>
<thead>
<tr>
<th>Term</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>edf. See Empirical distribution function (edf)</td>
<td>359</td>
</tr>
<tr>
<td>Effective dimension reduction (EDR)</td>
<td>359</td>
</tr>
<tr>
<td>direction, 321, 329, 331</td>
<td></td>
</tr>
<tr>
<td>space, 321</td>
<td></td>
</tr>
<tr>
<td>Efficient portfolio, 310</td>
<td></td>
</tr>
<tr>
<td>Eigenvalue, xxi, 22, 42, 168, 181, 183, 192, 247, 260, 263, 285</td>
<td></td>
</tr>
<tr>
<td>Eigenvector, xxi, 168, 181, 192, 247, 260, 262, 285</td>
<td></td>
</tr>
<tr>
<td>Elastic net, 158, 159</td>
<td></td>
</tr>
<tr>
<td>Ellipsoid, 23</td>
<td></td>
</tr>
<tr>
<td>Empirical distribution function (edf), xxi</td>
<td></td>
</tr>
<tr>
<td>Empirical moments, xxi</td>
<td></td>
</tr>
<tr>
<td>Entropy index, 321</td>
<td></td>
</tr>
<tr>
<td>Error of misclassification, 245</td>
<td></td>
</tr>
<tr>
<td>Estimator, xxi</td>
<td></td>
</tr>
<tr>
<td>minimum variance unbiased, 97</td>
<td></td>
</tr>
<tr>
<td>Euclidean distance, 227, 234, 239, 289, 290</td>
<td></td>
</tr>
<tr>
<td>Expected value, xxi, 43</td>
<td></td>
</tr>
<tr>
<td>conditional, xvi</td>
<td></td>
</tr>
<tr>
<td>Exploratory projection pursuit, 320, 327</td>
<td></td>
</tr>
<tr>
<td>Exponential distribution, 69</td>
<td></td>
</tr>
<tr>
<td>Factor, 167, 181, 301</td>
<td></td>
</tr>
<tr>
<td>analysis, 206</td>
<td></td>
</tr>
<tr>
<td>estimation, 206</td>
<td></td>
</tr>
<tr>
<td>factor scores, 207</td>
<td></td>
</tr>
<tr>
<td>oblique rotation, 206</td>
<td></td>
</tr>
<tr>
<td>promax rotation, 206</td>
<td></td>
</tr>
<tr>
<td>rotation, 206</td>
<td></td>
</tr>
<tr>
<td>strategy, 207</td>
<td></td>
</tr>
<tr>
<td>testing, 213</td>
<td></td>
</tr>
<tr>
<td>varimax rotation, 206</td>
<td></td>
</tr>
<tr>
<td>loadings, 205</td>
<td></td>
</tr>
<tr>
<td>scores, 207</td>
<td></td>
</tr>
<tr>
<td>estimation, 216</td>
<td></td>
</tr>
<tr>
<td>Factorial axis, 169</td>
<td></td>
</tr>
<tr>
<td>Factorial representation, 171</td>
<td></td>
</tr>
<tr>
<td>Factorial technique, 167</td>
<td></td>
</tr>
<tr>
<td>Factorial variable, 168, 171, 172</td>
<td></td>
</tr>
<tr>
<td>$F$-distribution, xviii</td>
<td></td>
</tr>
<tr>
<td>quantile, xviii</td>
<td></td>
</tr>
<tr>
<td>Feature space, 324</td>
<td></td>
</tr>
<tr>
<td>Fisher information matrix, 90, 96, 99</td>
<td></td>
</tr>
<tr>
<td>Fisher’s LDA, 245, 246, 255–257, 327</td>
<td></td>
</tr>
<tr>
<td>Fisher’s Z-transformation, 29, 34</td>
<td></td>
</tr>
<tr>
<td>Fitted values, 142, 148, 157</td>
<td></td>
</tr>
<tr>
<td>Five number summary, 6</td>
<td></td>
</tr>
<tr>
<td>Flury–Chernoff faces, 3, 7</td>
<td></td>
</tr>
<tr>
<td>Friedman–Tukey index, 320, 327</td>
<td></td>
</tr>
<tr>
<td>Gaussian distribution, xviii, xxiii</td>
<td></td>
</tr>
<tr>
<td>Generalized cross-validation, 161</td>
<td></td>
</tr>
<tr>
<td>Generalized linear model GLM, 145</td>
<td></td>
</tr>
<tr>
<td>GGobi, viii</td>
<td></td>
</tr>
<tr>
<td>Gini criterion, 323</td>
<td></td>
</tr>
<tr>
<td>glmnet, vii</td>
<td></td>
</tr>
<tr>
<td>Gradient, xxi</td>
<td></td>
</tr>
<tr>
<td>Gradient vector, 23</td>
<td></td>
</tr>
<tr>
<td>Graphical techniques, 3</td>
<td></td>
</tr>
<tr>
<td>Guided tour, 321</td>
<td></td>
</tr>
<tr>
<td>Halfspace depth, 320</td>
<td></td>
</tr>
<tr>
<td>Hat matrix, 148</td>
<td></td>
</tr>
<tr>
<td>Helmert contrasts, 146</td>
<td></td>
</tr>
<tr>
<td>Hessian matrix, xxii</td>
<td></td>
</tr>
<tr>
<td>Heteroscedasticity, 126</td>
<td></td>
</tr>
<tr>
<td>Hexagon plot, 3</td>
<td></td>
</tr>
<tr>
<td>hexbin, viii</td>
<td></td>
</tr>
<tr>
<td>Hierarchical clustering, 225</td>
<td></td>
</tr>
<tr>
<td>Histogram, 3, 7, 12</td>
<td></td>
</tr>
<tr>
<td>Horizontal profiles, 132, 134, 135</td>
<td></td>
</tr>
<tr>
<td>Hotelling distribution, xviii, 72, 108, 112, 126</td>
<td></td>
</tr>
<tr>
<td>quantile, xviii</td>
<td></td>
</tr>
<tr>
<td>Hypothesis testing, 103</td>
<td></td>
</tr>
<tr>
<td>Idempotent matrix, xxii, 23, 74</td>
<td></td>
</tr>
<tr>
<td>Impurity measure, 322</td>
<td></td>
</tr>
<tr>
<td>Independence, 29, 88, 260, 266</td>
<td></td>
</tr>
<tr>
<td>Indicator, xv</td>
<td></td>
</tr>
<tr>
<td>Inertia, 167, 168, 171, 173, 179, 181, 226, 227, 229</td>
<td></td>
</tr>
<tr>
<td>Intercluster distance, 226</td>
<td></td>
</tr>
<tr>
<td>Inverse matrix, 22, 24</td>
<td></td>
</tr>
<tr>
<td>Iso-distance curve, 50</td>
<td></td>
</tr>
<tr>
<td>Jacobian, xxii, 47, 61, 70</td>
<td></td>
</tr>
<tr>
<td>Joint distribution, 75, 79</td>
<td></td>
</tr>
<tr>
<td>Jones–Sibson index, 321</td>
<td></td>
</tr>
<tr>
<td>k-means inverse regression (KIR), 322</td>
<td></td>
</tr>
<tr>
<td>Kernel density estimator, 3, 7, 12, 320, 327</td>
<td></td>
</tr>
<tr>
<td>kernlab, viii</td>
<td></td>
</tr>
<tr>
<td>KernSmooth, viii</td>
<td></td>
</tr>
<tr>
<td>Kronecker product, xv</td>
<td></td>
</tr>
<tr>
<td>Jacobian, xxii, 47, 61, 70</td>
<td></td>
</tr>
<tr>
<td>Joint distribution, 75, 79</td>
<td></td>
</tr>
<tr>
<td>Jones–Sibson index, 321</td>
<td></td>
</tr>
<tr>
<td>k-means inverse regression (KIR), 322</td>
<td></td>
</tr>
<tr>
<td>Kernel density estimator, 3, 7, 12, 320, 327</td>
<td></td>
</tr>
<tr>
<td>kernlab, viii</td>
<td></td>
</tr>
<tr>
<td>KernSmooth, viii</td>
<td></td>
</tr>
<tr>
<td>Kronecker product, xv</td>
<td></td>
</tr>
<tr>
<td>$L_1$-distance, 233, 235</td>
<td></td>
</tr>
<tr>
<td>$L_2$-distance, 239</td>
<td></td>
</tr>
<tr>
<td>Lasso, 158, 159</td>
<td></td>
</tr>
<tr>
<td>$p$-value, 163</td>
<td></td>
</tr>
<tr>
<td>lasso2, viii, 161</td>
<td></td>
</tr>
<tr>
<td>Least squares, 31, 40, 41</td>
<td></td>
</tr>
<tr>
<td>constrained, 41</td>
<td></td>
</tr>
<tr>
<td>criterion, 323</td>
<td></td>
</tr>
<tr>
<td>Leverage, 149</td>
<td></td>
</tr>
</tbody>
</table>
### Index

<table>
<thead>
<tr>
<th>Term</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>xxii</td>
</tr>
<tr>
<td>function</td>
<td>89</td>
</tr>
<tr>
<td>ratio test</td>
<td>103, 104, 145</td>
</tr>
<tr>
<td>Linear approximation</td>
<td>85</td>
</tr>
<tr>
<td>Linear constraint</td>
<td>41, 132, 144</td>
</tr>
<tr>
<td>Linear dependence</td>
<td>xxii</td>
</tr>
<tr>
<td>Linear discrimination analysis</td>
<td>See Fisher’s LDA</td>
</tr>
<tr>
<td>Linear hypothesis</td>
<td>126</td>
</tr>
<tr>
<td>Linear model</td>
<td>40, 41, 45, 130, 131, 141, 301, 303</td>
</tr>
<tr>
<td>Linear predictor</td>
<td>145</td>
</tr>
<tr>
<td>Linear regression</td>
<td>130, 131, 142</td>
</tr>
<tr>
<td>Linear space</td>
<td>xviii</td>
</tr>
<tr>
<td>Linear transformation</td>
<td>45, 50, 52, 70, 72, 79–81, 83, 87</td>
</tr>
<tr>
<td>Link function</td>
<td>145</td>
</tr>
<tr>
<td>Liters per 100 km</td>
<td>51</td>
</tr>
<tr>
<td>Liu depth</td>
<td>320</td>
</tr>
<tr>
<td>locpol</td>
<td>viii</td>
</tr>
<tr>
<td>Log odds</td>
<td>144</td>
</tr>
<tr>
<td>Log-likelihood function</td>
<td>xxii, 89, 104, 108</td>
</tr>
<tr>
<td>Log-linear model</td>
<td>144</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>144</td>
</tr>
<tr>
<td>Logit</td>
<td>145</td>
</tr>
<tr>
<td>L_2-distance</td>
<td>234</td>
</tr>
<tr>
<td>Mahalanobis distance</td>
<td>247</td>
</tr>
<tr>
<td>Mahalanobis transformation</td>
<td>44, 192, 247, 320, 327</td>
</tr>
<tr>
<td>Manhattan distance</td>
<td>234</td>
</tr>
<tr>
<td>Marginal distribution</td>
<td>xxii, 46, 69, 75</td>
</tr>
<tr>
<td>Marginal moments</td>
<td>xxii</td>
</tr>
<tr>
<td>Marginal pdf</td>
<td>62</td>
</tr>
<tr>
<td>Marketing</td>
<td>301</td>
</tr>
<tr>
<td>MASS</td>
<td>vii, 158, 263</td>
</tr>
<tr>
<td>MATLAB</td>
<td>viii</td>
</tr>
<tr>
<td>Matrix</td>
<td></td>
</tr>
<tr>
<td>adjoint</td>
<td>xix, 22</td>
</tr>
<tr>
<td>centering</td>
<td>xvii, 36, 74</td>
</tr>
<tr>
<td>cofactor</td>
<td>xx</td>
</tr>
<tr>
<td>contrast</td>
<td>132</td>
</tr>
<tr>
<td>correlation</td>
<td>29, 36</td>
</tr>
<tr>
<td>covariance</td>
<td>xvi, 37, 42, 43, 55, 72</td>
</tr>
<tr>
<td>diagonal</td>
<td>98</td>
</tr>
<tr>
<td>determinant of</td>
<td>xviii</td>
</tr>
<tr>
<td>diagonal of</td>
<td>xviii</td>
</tr>
<tr>
<td>distance</td>
<td>289</td>
</tr>
<tr>
<td>Fisher information</td>
<td>90, 96, 99</td>
</tr>
<tr>
<td>gradient</td>
<td>xxi</td>
</tr>
<tr>
<td>Hessian</td>
<td>xxii</td>
</tr>
<tr>
<td>idempotent</td>
<td>xxii, 23, 74</td>
</tr>
<tr>
<td>inverse</td>
<td>22</td>
</tr>
<tr>
<td>Jacobian</td>
<td>xxii</td>
</tr>
<tr>
<td>orthogonal</td>
<td>xxiii, 211</td>
</tr>
<tr>
<td>partitioned</td>
<td>xxiii, 24, 26</td>
</tr>
<tr>
<td>projection</td>
<td>23</td>
</tr>
<tr>
<td>rank of</td>
<td>xviii</td>
</tr>
<tr>
<td>rotation</td>
<td>211</td>
</tr>
<tr>
<td>scatterplot</td>
<td>3, 15</td>
</tr>
<tr>
<td>trace</td>
<td>xviii</td>
</tr>
<tr>
<td>variance</td>
<td>42</td>
</tr>
<tr>
<td>Maximum likelihood estimator (MLE)</td>
<td>89, 90, 95, 97</td>
</tr>
<tr>
<td>MDS</td>
<td>See Multidimensional scaling (MDS)</td>
</tr>
<tr>
<td>Mean</td>
<td>xvi, xxii, xxiii</td>
</tr>
<tr>
<td>Mean squared error (MSE)</td>
<td>xxii, 69, 142</td>
</tr>
<tr>
<td>Mean-variance efficient portfolio</td>
<td>310</td>
</tr>
<tr>
<td>Mean-variance optimization</td>
<td>309</td>
</tr>
<tr>
<td>Median</td>
<td>xxii, 324</td>
</tr>
<tr>
<td>multivariate</td>
<td>319</td>
</tr>
<tr>
<td>linkage</td>
<td>226</td>
</tr>
<tr>
<td>Miles per gallon</td>
<td>51</td>
</tr>
<tr>
<td>Minimum variance unbiased estimator</td>
<td>97</td>
</tr>
<tr>
<td>Minor</td>
<td>See Cofactor</td>
</tr>
<tr>
<td>Misclassification rate</td>
<td>258</td>
</tr>
<tr>
<td>ML discrimination rule</td>
<td>245–247, 249, 251, 255, 256</td>
</tr>
<tr>
<td>MLE</td>
<td>See Maximum likelihood estimator (MLE)</td>
</tr>
<tr>
<td>Moments</td>
<td>xvi, xxii</td>
</tr>
<tr>
<td>empirical</td>
<td>xxi</td>
</tr>
<tr>
<td>marginal</td>
<td>xxii</td>
</tr>
<tr>
<td>MSE</td>
<td>See Mean squared error (MSE)</td>
</tr>
<tr>
<td>Multicollinearity</td>
<td>157</td>
</tr>
<tr>
<td>Multidimensional scaling (MDS)</td>
<td>289</td>
</tr>
<tr>
<td>metric</td>
<td>289, 291</td>
</tr>
<tr>
<td>nonmetric</td>
<td>289, 290, 293</td>
</tr>
<tr>
<td>Multinomial distribution</td>
<td>105</td>
</tr>
<tr>
<td>Multinormal distribution</td>
<td>xxiii</td>
</tr>
<tr>
<td>Multivariate median</td>
<td>320</td>
</tr>
<tr>
<td>Multivariate normal distribution</td>
<td>71</td>
</tr>
<tr>
<td>mvpart</td>
<td>viii</td>
</tr>
<tr>
<td>Nonlinear transformation</td>
<td>51, 60, 94</td>
</tr>
<tr>
<td>Normal distribution</td>
<td>xviii, xxiii</td>
</tr>
<tr>
<td>Normalized principal component analysis (NCPA)</td>
<td>192, 197, 200, 201</td>
</tr>
<tr>
<td>Null deviance</td>
<td>155</td>
</tr>
<tr>
<td>Observation</td>
<td>xvii</td>
</tr>
<tr>
<td>Odds</td>
<td>144</td>
</tr>
<tr>
<td>Order statistic</td>
<td>xvii</td>
</tr>
<tr>
<td>Orthogonal complement</td>
<td>xxiii</td>
</tr>
</tbody>
</table>
Index

Orthogonal factor model, 207
Orthogonal matrix, xxiii
Orthogonal polynomials, 146
Outlier, 3, 33

Parallel coordinate plot, 11, 12
Parallel profiles, 132, 135
Part-worth, 301, 303, 305, 307, 308
Partitioned covariance matrix, 71
Partitioned matrix, xxiii, 24, 26
PAV algorithm. See Pool-adjacent-violators (PAV) algorithm
PC. See Principal component (PC)
pdf. See Probability density function (pdf)
PhD. See Principal Hessian directions (PhD)
Polynomial contrasts, 146
Pool-adjacent-violators (PAV) algorithm, 301
Population, 245
Portfolio optimization, 309
Prediction interval, 143
Preference, 301, 307
Principal component (PC), 183, 191, 192, 243, 259, 278, 325
correlation, 184
expected value, 184
explained variance, 187
normalized, 192, 197, 200, 201, 278
screeplot, 185
testing, 187
variance, 184
Principal component analysis, 184
Principal factors, 206–208
Principal Hessian directions (PhD), 322
Prior probability, 246
Probability density function (pdf), xvi, 43, 48
conditional, xvi, 62
joint, xvi
marginal, xvi, 62
Probability of misclassification, 257
Profile analysis, 132, 135
Profile method, 301, 307
Projection, 169
Fisher’s LDA, 246
matrix, 23
pursuit, 319, 320, 327
vector, 321
Promax rotation, 206
Proximity, 229
p-value, xxiii

quadprog, viii
Quantile, xxiii

Quantitative finance, 309
Quantlet download center, viii, 343

R, vii, 263
Random sample, 89
Random variable, xv, xxiii
Random vector, xv, xxiii, 43
Rank, xviii
Regression
diagnostics, 151
line, 30, 33, 34, 37
tree (see CART)
Rejection region, 104
Residuals, 148
deviance, 155
standardized, 149
Response, 145
rggobi, viii
Ridge regression, 158
Risk management, 309
Rotation matrix, 211
Row coordinates, 263
Row factor, 263, 264
contributions to variance, 264
Row frequency, 259

Sample, xvii
Scatterplot, xxiii, 16, 28, 32
matrix, 3, 15
3D, 16
scatterplot3d, viii
Score function, 89, 97, 99–101
Screeplot, 185, 191
Semi-invariants, xvi
Separating hyperplane, 323, 337
Shepard–Kruskal algorithm, 290, 294
Shrinkage, 158
Simplicial depth, 319, 320, 324
Simultaneous testing, 144
Single linkage, 226, 230, 233
Singular value decomposition (SVD), xxiv, 169, 260, 263, 281, 284, 285
Sliced inverse regression (SIR) algorithm, 158, 319, 321, 322, 329, 331
Sliced inverse regression II (SIR II) algorithm, 322, 331
SMSdata, viii, 343
Specific factors, 205
Specific variance, 206
Spectral decomposition, xxiv, 22, 42, 44, 74, 170, 181, 183, 191, 290
Spiral, 338
<table>
<thead>
<tr>
<th>Standard error, 142</th>
<th>Two-factor method, 301</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardization, 36</td>
<td>Twoing criterion, 301</td>
</tr>
<tr>
<td>Statistical learning theory, 323</td>
<td>Type I error, 120</td>
</tr>
<tr>
<td>Stepwise regression, 159</td>
<td></td>
</tr>
<tr>
<td>Stimulus, 301</td>
<td></td>
</tr>
<tr>
<td>STRESS, 290, 303</td>
<td></td>
</tr>
<tr>
<td>Subspace, xxiv</td>
<td></td>
</tr>
<tr>
<td>Sum of squares, 30, 32</td>
<td></td>
</tr>
<tr>
<td>Support vector machine (SVM), 323</td>
<td></td>
</tr>
<tr>
<td>SVD. See Singular value decomposition (SVD)</td>
<td></td>
</tr>
<tr>
<td>Taylor expansion, xxiv, 125</td>
<td></td>
</tr>
<tr>
<td>$t$-distribution quantile, xviii</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>covariance matrix, 118, 120, 129</td>
<td></td>
</tr>
<tr>
<td>equality of means, 111</td>
<td></td>
</tr>
<tr>
<td>expected value, 118</td>
<td></td>
</tr>
<tr>
<td>independence, 123, 127</td>
<td></td>
</tr>
<tr>
<td>likelihood ratio, 104, 145</td>
<td></td>
</tr>
<tr>
<td>linear model, 130</td>
<td></td>
</tr>
<tr>
<td>mean vector, 113–116, 121, 123, 127</td>
<td></td>
</tr>
<tr>
<td>number of factors, 213</td>
<td></td>
</tr>
<tr>
<td>principal component, 187</td>
<td></td>
</tr>
<tr>
<td>two-sample, 123</td>
<td></td>
</tr>
<tr>
<td>Theorem</td>
<td></td>
</tr>
<tr>
<td>binomial inverse, 25</td>
<td></td>
</tr>
<tr>
<td>central limit (see CLT)</td>
<td></td>
</tr>
<tr>
<td>Cramer–Rao, 90</td>
<td></td>
</tr>
<tr>
<td>Wilks’, 104</td>
<td></td>
</tr>
<tr>
<td>Trace, xviii, 36</td>
<td></td>
</tr>
<tr>
<td>Transformation of statistics, 51, 94</td>
<td></td>
</tr>
<tr>
<td>tseries, viii</td>
<td></td>
</tr>
<tr>
<td>Tukey depth, 320</td>
<td></td>
</tr>
<tr>
<td>Variable selection, 158</td>
<td></td>
</tr>
<tr>
<td>Variance, xvi, 54</td>
<td></td>
</tr>
<tr>
<td>conditional, xvi, 47</td>
<td></td>
</tr>
<tr>
<td>efficient portfolio, 310</td>
<td></td>
</tr>
<tr>
<td>empirical, xvii</td>
<td></td>
</tr>
<tr>
<td>matrix, 42</td>
<td></td>
</tr>
<tr>
<td>Variance inflation factor (VIF), 157, 161</td>
<td></td>
</tr>
<tr>
<td>Varimax rotation, 206</td>
<td></td>
</tr>
<tr>
<td>Vector</td>
<td></td>
</tr>
<tr>
<td>gradient, xxi, 23</td>
<td></td>
</tr>
<tr>
<td>VIF () , 161</td>
<td></td>
</tr>
<tr>
<td>Ward method, 226, 227, 229, 230, 235</td>
<td></td>
</tr>
<tr>
<td>White noise analysis, 321</td>
<td></td>
</tr>
<tr>
<td>Wilks’ theorem, 104</td>
<td></td>
</tr>
<tr>
<td>Wishart distribution, 72, 74, 133</td>
<td></td>
</tr>
<tr>
<td>Within variance, 246</td>
<td></td>
</tr>
<tr>
<td>Within-group sum of squares, 246, 252</td>
<td></td>
</tr>
<tr>
<td>zoo, viii</td>
<td></td>
</tr>
</tbody>
</table>