

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

- X_3 : C nondepreciation of value
- X_4 : D price, mark 1 for very cheap cars
- X_5 : E design
- X_6 : F sporty car
- X_7 : G safety
- X_8 : H easy handling

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

- X_6 : F industrial techniques
- X_7 : G economic techniques
- X_8 : H computer 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 : rnnr 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 :	WaBr	Walloon Brabant
X_2 :	Brar	Brussels area
X_3 :	Antw	Antwerp
X_4 :	FlBr	Flemish Brabant
X_5 :	OcFl	Occidental Flanders
X_6 :	OrFl	Oriental Flanders
X_7 :	Hain	Hainaut
X_8 :	Lièg	Liège
X_9 :	Limb	Limburg
X_{10} :	Luxe	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 :	8 AM
X_2 :	11 AM
X_3 :	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 ($100 \cdot 24 \text{ h} = 2,400 \text{ h}$ 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

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)

Division Numbers		Region Numbers	
New England	1	Northeast	1
Mid-Atlantic	2	Midwest	2
E N Central	3	South	3
W N Central	4	West	4
S Atlantic	5		
E S Central	6		
W S Central	7		
Mountain	8		
Pacific	9		

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:

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

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).

A.21 WAIS Data

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

- Andrews, D. (1972). Plots of high-dimensional data. *Biometrics*, 28, 125–136.
- Bartlett, M. S. (1954). A note on multiplying factors for various chi-squared approximations. *Journal of the Royal Statistical Society, Series B*, 16, 296–298.
- Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language. A programming environment for data analysis and graphics*. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics*. New York: Wiley.
- Berndt, E. R. (1990). *The practice of econometrics: Classic and contemporary*. Reading: Addison-Wesley.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Bouroche, J.-M., & Saporta, G. (1980). *L'analyse des données*. Paris: Presses Universitaires de France.
- Breiman, L. (1973). *Statistics: With a view towards application*. Boston: Houghton Mifflin Company.
- Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. J. (1984). *Classification and regression trees*. Belmont: Wadsworth.
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data*, Springer Series in Statistics. Heidelberg: Springer.
- Cabrera, J. L. O. (2012). *locpol: Kernel local polynomial regression*. R package version 0.6-0.
- Carr, D., Lewin-Koh, N., & Maechler, M. (2011). *hexbin: Hexagonal binning routines*. R package version 1.26.0.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Boston: Duxbury.
- Chambers, J. M., & Hastie, T. J. (1992). *Statistical models in S*. Boca Raton: Chapman & Hall/CRC.
- Chernoff, H. (1973). Using faces to represent points in k -dimensional space graphically. *Journal of the American Statistical Association*, 68, 361–368.
- Cook, D., Buja, A., & Cabrera, J. (1993). Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics*, 2(3), 225–250.
- Cook, R. D. (1998). *Regression graphics*. New York: Wiley.
- Cook, R. D., & Weisberg, S. (1991). Comment on sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414), 328–332.
- Duan, N., & Li, K.-C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*, 19(2), 505–530.

- Eddy, W. (1982). Convex hull peeling, in H. Caussinus, P. Ettinger, & R. Tomassone (Eds.), *COMPSTAT 1982 5th Symposium held at Toulouse 1982* (pp. 42–47). Heidelberg: Physica-Verlag.
- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: models, methods and applications*. New York: Springer.
- Feller, W. (1966). *An introduction to probability theory and its application* (Vol. 2). New York: Wiley.
- Flury, B., & Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association*, 76, 757–765.
- Flury, B., & Riedwyl, H. (1988). *Multivariate statistics, a practical approach*. London: Chapman and Hall.
- Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.
- Franke, J., Härdle, W., & Hafner, C. (2011). *Statistics of financial markets: An introduction* (3rd ed.). Berlin: Springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
- Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C23, 881–890.
- Genest, M., Masse, J.-C., & Plante, J.-F. (2012). *depth: Depth functions tools for multivariate analysis*. R package version 2.0-0.
- Hall, P., & Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, 21(2), 867–889.
- Härdle, W., Moro, R., & Schäfer, D. (2005). Predicting bankruptcy with support vector machines, in P. Čížek, W. Härdle, & R. Weron (Eds.), *Statistical Tools for Finance and Insurance* (pp. 225–248). Berlin: Springer.
- Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Berlin: Springer.
- Härdle, W., & Simar, L. (2015). *Applied multivariate statistical analysis* (4th ed.). Berlin: Springer.
- Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.
- Harville, D. A. (2001). *Matrix algebra: Exercises and solutions*. New York: Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, Springer Series in Statistics (2nd ed.). New York: Springer.
- Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1), 65–70.
- Hlubinka, D., Kotík, L., & Vencálek, O. (2010). Weighted halfspace depth. *Kybernetika*, 46(1), 125–148.
- Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Hui, G., & Lindsay, B. G. (2010). Projection pursuit via white noise matrices. *Sankhya B*, 72(2), 123–153. With a discussion by Surajit Ray and a rejoinder by the authors.
- Johnson, R. A., & Wichern, D. W. (1998). *Applied Multivariate Analysis* (4th ed.). Englewood Cliffs: Prentice Hall.
- Jones, M. C., & Sibson, R. (1987). What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society, Series A*, 150(1), 1–36.
- Kaiser, H. F. (1985). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9), 1–20.
- Knight, K., & Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics*, 28(5), 1356–1378.
- Kötter, T. (1996). Entwicklung statistischer Software, Ph.D. thesis, Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin.
- Kruskal, J. B. (1965). Analysis of factorial experiments by estimating a monotone transformation of data. *Journal of the Royal Statistical Society, Series B*, 27, 251–263.

- Lang, D. T., Swayne, D., Wickham, H., & Lawrence, M. (2012). *rggobi: Interface between R and GGobi*. R package version 2.1.19.
- Lebart, L., Morineau, A., & F  nelon, J. P. (1982). *Traitement des Donn  es Statistiques: M  thodes et programmes*. Paris: Dunod.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414), 316–342.
- Ligges, U., & M  chler, M. (2003). Scatterplot3d—an R package for visualizing multivariate data. *Journal of Statistical Software*, 8(11), 1–20.
- Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Sciences USA*, 85(6), 1732–1734.
- Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1), 405–414.
- Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2013). A significance test for the lasso. arXiv preprint arXiv:1301.7161.
- Lokhorst, J., Venables, B., Turlach, B., & Maechler, M. (2013). *lasso2: L1 constrained estimation aka lasso*. R package version 1.2-16.
- L  tkepohl, H. (1996). *Handbook of matrices*. Chichester: Wiley.
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. Duluth/London: Academic.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, Monographs on Statistics and Applied Probability (2nd ed., Vol. 37). London: Chapman and Hall.
- Mizera, I., & M  ller, C. H. (2004). Location-scale depth. *Journal of the American Statistical Association*, 99(468), 949–989. With comments and a rejoinder by the authors.
- Morrison, D. F. (1990). *Multivariate statistical methods*. New York: McGraw-Hill.
- Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, 20(3), 1–13.
- Neter, J., Wasserman, W., Kutner, M. H., & Wasserman, W. (1996). *Applied linear statistical models*. (4 ed.). Chicago: Irwin.
- Olkin, I., & Veath, M. (1980). Maximum likelihood estimation in a two-way analysis with correlated errors in one classification. *Biometrika*, 68, 653–660.
- Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2), 319–337.
- Puntenan, S., Sty  n, G. P. H., & Isotalo, J. (2011). *Matrix tricks for linear statistical models*. Heidelberg: Springer.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.
- Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, 53(4), 382–387.
- Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425), 141–148.
- Searle, S. R. (1982). *Matrix algebra useful for statistics*. Chichester: Wiley.
- Seber, G. A. F. (2008). *A matrix handbook for statisticians*. Hoboken, NJ: Wiley.
- SEC (2004). Archive of historical documents, Securities and Exchange Commission. www.sec.gov/cgi-bin/srch-edgar.
- Serfling, R. J. (2002). *Approximation theorems of mathematical statistics*. New York: Wiley.
- Setodji, C. M., & Cook, R. D. (2004). K-means inverse regression. *Technometrics*, 46(4), 421–429.
- Sobel, R. (1988). *Panic on Wall Street: A classic history of America's financial disasters with a new exploration of the crash of 1987*. New York: Truman Talley Books/Dutton.
- Swayne, D. F., Lang, D. T., Buja, A., & Cook, D. (2003). GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics and Data Analysis*, 43(4), 423–444.
- Therneau, T. M., Atkinson, B., Ripley, B., Oksanen, J., & De  th, G. (2012). *mvpart: Multivariate partitioning*. R package version 1.6-0.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.

- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society, Series B*, 73(3), 273–282.
- Trapletti, A., & Hornik, K. (2012). *tseries: Time series analysis and computational finance*. R package version 0.10-30.
- Tukey, J. W. (1975). Mathematics and the picturing of data, in *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974)* (Vol. 2, pp. 523–531). Montreal, QC: Canadian Mathematical Congress.
- Turlach, B. A., & Weingessel, A. (2011). *quadprog: Functions to solve quadratic programming problems*. R package version 1.5-4.
- Vapnik, V. N. (2000). *The nature of statistical learning theory* (2nd ed.). New York: Springer.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
- Volle, V. (1985). *Analyse des Données*. Paris: Economica.
- Wand, M. (2012). *KernSmooth: Functions for kernel smoothing for Wand & Jones (1995)*. R package version 2.23-8.
- Wang, Y. (2012). Model selection, in J. E. Gentle, W. K. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics* (2nd ed., pp. 469–498). Berlin: Springer.
- Ward, J. H. (1963). Hierarchical grouping methods to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Weisberg, S. (2002). Dimension reduction regression in R. *Journal of Statistical Software*, 7(1), 1–22.
- Wolf, P. (2012). *aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions*. R package version 1.2.7.
- Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6), 1–27.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320.
- Zuo, Y., & Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2), 461–482.
- Zvára. (2008). *Regression (in Czech)*, Prague: Matfyzpress.

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)
- ap1pack, 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

- ca, viii, 263
- Canonical correlation analysis (CCA)
 - analysis, 281, 283, 287
 - coefficient, 281, 282
 - variables, 281, 282, 284
 - 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
- χ^2 distance, 239
- χ^2 distribution, xviii
 - quantile, xviii
- χ^2 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

- 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
 - χ^2 , 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

- edf. *See* Empirical distribution function (edf)
- Effective dimension reduction (EDR)
 - direction, 321, 329, 331
 - space, 321
- Efficient portfolio, 310
- Eigenvalue, xxi, 22, 42, 168, 181, 183, 192, 247, 260, 263, 285
- Eigenvector, xxi, 168, 181, 192, 247, 260, 262, 285
- Elastic net, 158, 159
- Ellipsoid, 23
- Empirical distribution function (edf), xxi
- Empirical moments, xxi
- Entropy index, 321
- Error of misclassification, 245
- Estimator, xxi
 - minimum variance unbiased, 97
- Euclidean distance, 227, 234, 239, 289, 290
- Expected value, xxi, 43
 - conditional, xvi
- Exploratory projection pursuit, 320, 327
- Exponential distribution, 69

- Factor, 167, 181, 301
 - analysis, 206
 - estimation, 206
 - factor scores, 207
 - oblique rotation, 206
 - promax rotation, 206
 - rotation, 206
 - strategy, 207
 - testing, 213
 - varimax rotation, 206
 - loadings, 205
 - scores, 207
 - estimation, 216
- Factorial axis, 169
- Factorial representation, 171
- Factorial technique, 167
- Factorial variable, 168, 171, 172
- F*-distribution, xviii
 - quantile, xviii
- Feature space, 324
- Fisher information matrix, 90, 96, 99
- Fisher's LDA, 245, 246, 255–257, 327
- Fisher's *Z*-transformation, 29, 34
- Fitted values, 142, 148, 157
- Five number summary, 6
- Flury–Chernoff faces, 3, 7
- Friedman–Tukey index, 320, 327

- Gaussian distribution, xviii, xxiii
- Generalized cross-validation, 161
- Generalized linear model GLM, 145
- GGobi, viii, 321
- Gini criterion, 323
- glmnet, vii, 163
- Gradient, xxi
- Gradient vector, 23
- Graphical techniques, 3
- Guided tour, 321

- Halfspace depth, 320
- Hat matrix, 148
- Helmert contrasts, 146
- Hessian matrix, xxii
- Heteroscedasticity, 126
- Hexagon plot, 3
- hexbin, viii
- Hierarchical clustering, 225
- Histogram, 3, 7, 12
- Horizontal profiles, 132, 134, 135
- Hotelling distribution, xviii, 72, 108, 112, 126
 - quantile, xviii
- Hypothesis testing, 103

- Idempotent matrix, xxii, 23, 74
- Impurity measure, 322
- Independence, 29, 88, 260, 266
- Indicator, xv
- Inertia, 167, 168, 171, 173, 179, 181, 226, 227, 229
- Intercluster distance, 226
- Inverse matrix, 22, 24
- Iso-distance curve, 50

- Jacobian, xxii, 47, 61, 70
- Joint distribution, 75, 79
- Jones–Sibson index, 321

- k*-means inverse regression (KIR), 322
- Kernel density estimator, 3, 7, 12, 320, 327
- kernlab, viii
- KernSmooth, viii
- Kronecker product, xv

- L_1 -distance, 233, 235
- L_2 -distance, 239
- Lasso, 158, 159
 - p*-value, 163
- lasso2, viii, 161
- Least squares, 31, 40
 - constrained, 41
 - criterion, 323
- Leverage, 149

- Likelihood, xxii
 - function, 89
 - ratio test, 103, 104, 145
- Linear approximation, 85
- Linear constraint, 41, 132, 144
- Linear dependence, xxii
- Linear discrimination analysis. *See* Fisher's LDA
- Linear hypothesis, 126
- Linear model, 40, 41, 45, 130, 131, 141, 301, 303
- Linear predictor, 145
- Linear regression, 130, 131, 142
- Linear space, xviii
- Linear transformation, 45, 50, 52, 70, 72, 79–81, 83, 87
- Link function, 145
- Liters per 100 km, 51
- Liu depth, 320
- logcpol, viii
- Log odds, 144
- Log-likelihood function, xxii, 89, 104, 108
- Log-linear model, 144
- Logistic regression, 144
- Logit, 145
- L_r -distance, 234

- Mahalanobis distance, 247
- Mahalanobis transformation, 44, 192, 247, 320, 327
- Manhattan distance, 234
- Marginal distribution, xxii, 46, 69, 75
- Marginal moments, xxii
- Marginal pdf, 62
- Marketing, 301
- MASS, vii, 158, 263
- MATLAB, viii
- Matrix
 - adjoint, xix, 22
 - centering, xvii, 36, 74
 - cofactor, xx
 - contrast, 132
 - correlation, 29, 36
 - covariance, xvi, 37, 42, 43, 55, 72
 - diagonal, 98
 - determinant of, xviii
 - diagonal of, xviii
 - distance, 289
 - Fisher information, 90, 96, 99
 - gradient, xxi
 - Hessian, xxii
 - idempotent, xxii, 23, 74
 - inverse, 22
 - Jacobian, xxii
 - orthogonal, xxiii, 211
 - partitioned, xxiii, 24, 26
 - projection, 23
 - rank of, xviii
 - rotation, 211
 - scatterplot, 3, 15
 - trace, xviii
 - variance, 42
- Maximum likelihood estimator (MLE), 89, 90, 95, 97
- MDS. *See* Multidimensional scaling (MDS)
- Mean, xvi, xxi, xxii
- Mean squared error (MSE), xxii, 69, 142
- Mean-variance efficient portfolio, 310
- Mean-variance optimization, 309
- Median, xxii, 324
 - multivariate, 319
 - linkage, 226
- Miles per gallon, 51
- Minimum variance unbiased estimator, 97
- Minor. *See* Cofactor
- Misclassification rate, 258
- ML discrimination rule, 245–247, 249, 251, 255, 256
- MLE. *See* Maximum likelihood estimator (MLE)
- Moments, xvi, xxii
 - empirical, xxi
 - marginal, xxii
- MSE. *See* Mean squared error (MSE)
- Multicollinearity, 157
- Multidimensional scaling (MDS), 289
 - metric, 289, 291
 - nonmetric, 289, 290, 293
- Multinomial distribution, 105
- Multinormal distribution, xxiii
- Multivariate median, 320
- Multivariate normal distribution, 71
- mvpart, viii

- Nonlinear transformation, 51, 60, 94
- Normal distribution, xviii, xxiii
- Normalized principal component analysis (NCPA), 192, 197, 200, 201
- Null deviance, 155

- Observation, xvii
- Odds, 144
- Order statistic, xvii
- Orthogonal complement, xxiii

- 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, 302
- 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

- Standard error, 142
- Standardization, 36
- Statistical learning theory, 323
- Stepwise regression, 159
- Stimulus, 301
- STRESS, 290, 303
- Subspace, xxiv
- Sum of squares, 30, 32
- Support vector machine (SVM), 323
- SVD. *See* Singular value decomposition (SVD)

- Taylor expansion, xxiv, 125
- t*-distribution quantile, xviii
- Test
 - covariance matrix, 118, 120, 129
 - equality of means, 111
 - expected value, 118
 - independence, 123, 127
 - likelihood ratio, 104, 145
 - linear model, 130
 - mean vector, 113–116, 121, 123, 127
 - number of factors, 213
 - principal component, 187
 - two-sample, 123
- Theorem
 - binomial inverse, 25
 - central limit (*see* CLT)
 - Cramer–Rao, 90
 - Wilks', 104
- Trace, xviii, 36
- Transformation of statistics, 51, 94
- tseries*, viii
- Tukey depth, 320

- Two-factor method, 301
- Twoing criterion, 323
- Type I error, 120

- Unbiased estimator, 90
- Uniform distribution, 55, 90, 197, 200
- Utility, 306
- Utility attribute, 301

- Variable selection, 158
- Variance, xvi, 54
 - conditional, xvi, 47
 - efficient portfolio, 310
 - empirical, xvii
 - matrix, 42
- Variance inflation factor (VIF), 157, 161
- Varimax rotation, 206
- Vector
 - gradient, xxi, 23
- VIF (), 161

- Ward method, 226, 227, 229, 230, 235
- White noise analysis, 321
- Wilks' theorem, 104
- Wishart distribution, 72, 74, 133
- Within variance, 246
- Within-group sum of squares, 246, 252

- zoo, viii