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.

W.K. Härdle, Z. Hlávka, *Multivariate Statistics*,
DOI 10.1007/978-3-642-36005-3

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:	11AM
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 (100·24 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

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.

W.K. Härdle, Z. Hlávka, *Multivariate Statistics*,
DOI 10.1007/978-3-642-36005-3

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., & Fénelon, J. P. (1982). *Traitement des Donnés Statistiques: Méthodes 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., & Mächler, 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.

Lütkepohl, 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., & Müller, 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*. NewYork: 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.

Puntanen, S., Styan, 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., & Deáth, 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

W.K. Härdle, Z. Hlávka, *Multivariate Statistics*,
DOI 10.1007/978-3-642-36005-3