Modern Data Analysis in S-PLUS

References and On-line information

The main reference is Venables & Ripley (1997) and its on-line complements (especially those on statistics and programming) available at several sites via

`http://www.stats.ox.ac.uk/pub/MASS2/sites.html`

The scripts used in the demos are at

`http://www.stats.ox.ac.uk/pub/bdr/Washington98/`

On-line material

`XGobi` (Swayne *et al.*, 1998) is available for Unix as source files from `http://lib.stat.cmu.edu/general/XGobi`; a newer 'beta' version is available at

`http://www.research.att.com/~andreas/xgobi/`.

I have put up a version of `XGobi` for Windows 9x/NT at

`http://www.stats.ox.ac.uk/pub/SWin`.

All the S-PLUS libraries used are available for Windows at

`http://www.stats.ox.ac.uk/pub/SWin/`

Libraries `MASS` and `nnet` are part of the Venables & Ripley set, available for both Unix and Windows at `http://www.stats.ox.ac.uk/pub/MASS2/`

Library `KernSmooth` for Unix is at `http://www.biostat.harvard.edu/~mwand`

Library `locfit` for Unix and Windows is at

`http://cm.bell-labs.com/stat/project/locfit/`

and has on-line manuals (only).

Libraries `mda` and `pspline` are at `statlib`: `http://lib.stat.cmu.edu/S/`

Libraries `ppr` and `rpart` (Therneau & Atkinson, 1997) are at

`http://www.stats.ox.ac.uk/pub/S/`

References

This is an overview: in particular the books will give references to the primary sources.

Visualization

Ripley (1996, Chapter 9) gives an overview of multivariate visualization methods, including Sammon mapping (Sammon, 1969), multidimensional scaling (Shepard, 1962a,b; Kruskal, 1964a,b) and exploratory pursuit regression (Kruskal, 1969; Friedman & Tukey, 1974; Huber, 1985; Friedman, 1987; Jones & Sibson, 1987).

Density estimation and smoothing

Simonoff (1996) provides an excellent overview of density estimation and smoothing methods, with sections on computational resources including S-PLUS functions.

Bowman & Azzalini (1997) provide a gentle introduction to (kernel) density estimation and smoothing methods, and provide (on-line) an S-PLUS library with many examples. Wand & Jones (1995) give a more theoretical overview.

Local likelihood methods are discussed in Hjort & Jones (1996) and used in Loader (1996, 1997).

There are several closely-related proposals (see Simonoff, 1996, pp. 67–70, 90–92) for parametric fitting of smooth functions to log-densities. `logspline` implements that by Kooperberg & Stone (1992), although a later version described in Stone *et al.* (1997) is promised to replace it.

Beyond linear models

Hastie & Tibshirani (1990) is the most extensive reference for additive models, and covers `bruto`. `mars` is developed in Friedman (1991). Projection pursuit regression (Friedman & Stuetzle, 1981) and many other variants are described in Ripley (1996, Chapter 4).

Neural networks

There are very many books on neural networks, almost all algorithm-oriented. Bishop (1995) is perhaps the best of the traditional school. Statisticians will (I hope) find my work (Ripley, 1993, 1994, 1996) more accessible.

Tree-based methods

The main contribution by statisticians, Breiman *et al.* (1984), is still very relevant although the theory is unnecessarily complex. There are many other perspectives and contributions from computer science, engineering and machine learning. For example, Quinlan (1993) is a book documenting and listing a (widely-used) program, C4.5.

Ripley (1996, Chapter 7) gives an integrated approach with new, short, proofs.

LDA and extensions

The theory and formulae are all in Ripley (1996) in sections 2.4, 2.5 and 3.1.

Bibliography

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Bowman, A. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Monterey: Wadsworth and Brooks/Cole.

Friedman, J. H. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–266.

Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19**, 1–141.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–823.

Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **23**, 881–890.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.

Hjort, N. L. and Jones, M. C. (1996) Locally parametric nonparametric density estimation. *Annals of Statistics* **24**, 1619–1647.

Huber, P. J. (1985) Projection pursuit (with discussion). *Annals of Statistics* **13**, 435–525.

Jones, M. C. and Sibson, R. (1987) What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society A* **150**, 1–36.

Kooperberg, C. and Stone, C. J. (1992) Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**, 301–328.

Kruskal, J. B. (1964a) Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–29.

Kruskal, J. B. (1964b) Non-metric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115–129.

Kruskal, J. B. (1969) Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation*, eds R. C. Milton and J. A. Nelder, pp. 427–440. New York: Academic Press.

Loader, C. R. (1996) Local likelihood density estimation. *Annals of Statistics* **24**, 1602–1618.

Loader, C. R. (1997) Locfit: An introduction. *Statistical Computing and Graphics Newsletter* [Available from `http://cm.bell-labs.com/stat/project/locfit`].

Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Ripley, B. D. (1993) Statistical aspects of neural networks. In *Networks and Chaos – Statistical and Probabilistic Aspects*, eds O. E. Barndorff-Nielsen, J. L. Jensen and W. S. Kendall, pp. 40–123. London: Chapman and Hall.

Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society series B* **56**, 409–456.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Sammon, J. W. (1969) A non-linear mapping for data structure analysis. *IEEE Transactions on Computers* **C-18**, 401–409.

Shepard, R. N. (1962a) The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika* **27**, 125–139.

Shepard, R. N. (1962b) The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika* **27**, 219–246.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modelling. *Annals of Statistics* **25**, 1371–1470.

Swayne, D. F., Cook, D. and Buja, A. (1998) XGobi: Interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics* **7**, 113–130.

Therneau, T. M. and Atkinson, E. J. (1997) An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation. [Distributed in PostScript with the RPART package.].

Venables, W. N. and Ripley, B. D. (1997) *Modern Applied Statistics with S-PLUS*. Second Edition. New York: Springer-Verlag.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Chapman & Hall.