

References

- Abu-Mostafa, Y. (1995). Hints, *Neural Computation* **7**: 639–671.
- Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I. (1995). Fast discovery of association rules, *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge, MA.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, pp. 267–281.
- Allen, D. (1977). The relationship between variable selection and data augmentation and a method of prediction, *Technometrics* **16**: 125–7.
- Anderson, J. and Rosenfeld, E. (eds) (1988). *Neurocomputing: Foundations of Research*, MIT Press, Cambridge, MA.
- Barron, A. (1993). Universal approximation bounds for superpositions of a sigmoid function, *IEEE transactions on Information Theory* **39**: 930–945.
- Becker, R., Cleveland, W. and Shyu, M. (1996). The visual design and control of trellis display, *Journal of Computational and Graphical Statistics* **5**: 123–155.
- Bell, A. and Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution, *Neural Computation* **7**: 1129–1159.

- Bellman, R. E. (1961). *Adaptive Control Processes*, Princeton University Press.
- Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.
- Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error, *J. Amer. Statist. Assoc* **87**: 738–754.
- Breiman, L. (1996a). Bagging predictors, *Machine Learning* **26**: 123–140.
- Breiman, L. (1996b). Stacked regressions, *Machine Learning* **24**: 51–64.
- Breiman, L. (1998). Arcing classifiers (with discussion), *Annals of Statistics* **26**: 801–849.
- Breiman, L. (1999). Prediction games and arcing algorithms, *Neural Computation* pp. 1493–1517.
- Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression (with discussion), *J. Roy. Statist. Soc. B.* **59**: 3–37.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth.
- Breiman, L. and Ihaka, R. (1984). Nonlinear discriminant analysis via scaling and ACE, *Technical report*, Univ. of California, Berkeley.
- Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: the X-random case, *Intern. Statist. Rev* **60**: 291–319.
- Bruce, A. and Gao, H. (1996). *Applied Wavelet Analysis with S-PLUS*, Springer.
- Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**: 453–555.
- Buja, A., Swayne, D., Littman, M. and Dean, N. (1999). Xgvis: A system for multidimensional scaling and graph layout in any dimension, *Technical report*, AT&T Laboratories.
- Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition, *Knowledge Discovery and Data Mining* **2**(2): 121–167.
- Chambers, J. and Hastie, T. (1991). *Statistical Models in S*, Wadsworth/Brooks Cole, Pacific Grove, CA.
- Cherkassky, V. and Mulier, F. (1998). *Learning from Data*, Wiley, New York.

- Chui, C. (1992). *An Introduction to Wavelets*, Academic Press, London.
- Comon, P. (1994). Independent component analysis — a new concept?, *Signal Processing* **36**: 287–314.
- Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological* **45**: 311–354.
- Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *Proc. IEEE Trans. Inform. Theory* **IT-11**: 21–27.
- Cover, T. and Thomas, J. (1991). *Elements of Information Theory*, Wiley, New York.
- Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*, Chapman and Hall, London.
- Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised Edition)*, Wiley-Interscience, New York.
- Csiszar, I. and Tusnady, G. (1984). Information geometry and alternating minimization procedures, *Statistics & Decisions Supplement Issue* **1**: 205–237.
- Dasarathy, B. (1991). *Nearest Neighbor Pattern Classification Techniques*, IEEE Computer Society Press.
- Daubechies, I. (1992). *Ten Lectures in Wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA.
- de Boor, C. (1978). *A Practical Guide to Splines*, Springer-Verlag, New York.
- Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Statist. Soc. B.* **39**: 1–38.
- Devijver, P. and Kittler, J. (1982). *Pattern Recognition: a Statistical Approach*, Prentice-Hall, Englewood Cliffs, N.J.
- Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**: 425–455.
- Donoho, D., Johnstone, I., Kerkyachairan, G. and Picard, D. (1995). Wavelet shrinkage; asymptopia? (with discussion), *J. Royal. Statist. Soc.* **57**: 201–337.
- Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method, *Annals of Statistics* **19**: 505–530.

- Duchamp, T. and Stuetzle, W. (1996). Extremal properties of principal curves in the plane, *The Annals of Statistics* **24**: 1511–1520.
- Duda, R., Hart, P. and Stork, D. (2000). *Pattern Classification (Second Edition)*, Wiley, New York.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis, *J. Amer. Statist. Assoc.* **70**: 892–898.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics* **7**: 1–26.
- Efron, B. (1983). Estimating the error rate of a prediction rule: some improvements on cross-validation, *J. Amer. Statist. Assoc.* **78**: 316–331.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule?, *J. Amer. Statist. Assoc.* **81**: 461–70.
- Efron, B. and Tibshirani, R. (1991). Statistical analysis in the computer age, *Science* **253**: 390–395.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London.
- Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap: method, *J. Amer. Statist. Assoc.* **92**: 548–560.
- Evgeniou, T., Pontil, M. and Poggio, T. (2001). Regularization networks and support vector machines, *Advances in Computational Mathematics* (to appear).
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Eugen.* **7**: 179–188.
- Fix, E. and Hodges, J. (1951). Discriminatory analysis- nonparametric discrimination: Consistency properties, *Technical Report 21-49-004,4*, US Air Force, School of Aviation Medicine, Randolph Field, TX.
- Flury, B. (1990). Principal points, *Biometrika* **77**: 33–41.
- Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications (abstract), *Biometrics* **21**: 768–769.
- Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics* **35**(2): 109–148.

- Freund, Y. (1995). Boosting a weak learning algorithm by majority, *Information and Computation* **121**(2): 256–285.
- Freund, Y. and Schapire, R. (1996a). Experiments with a new boosting algorithm, *Machine Learning: Proceedings of the Thirteenth International Conference*, Morgan Kaufman, San Francisco, pp. 148–156.
- Freund, Y. and Schapire, R. (1996b). Game theory, on-line prediction and boosting, *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pp. 325–332.
- Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting, *Journal of Computer and System Sciences* **55**: 119–139.
- Friedman, J. (1987). Exploratory projection pursuit, *Journal of the American Statistical Association* **82**: 249–266.
- Friedman, J. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association* **84**: 165–175.
- Friedman, J. (1991). Multivariate adaptive regression splines (with discussion), *Annals of Statistics* **19**(1): 1–141.
- Friedman, J. (1994a). Flexible metric nearest-neighbour classification, *Technical report*, Stanford University.
- Friedman, J. (1994b). An overview of predictive learning and function approximation, in V. Cherkassky, J. Friedman and H. Wechsler (eds), *From Statistics to Neural Networks*, Vol. 136 of *NATO ISI Series F*, Springer Verlag, New York.
- Friedman, J. (1996). Another approach to polychotomous classification, *Technical report*, Stanford University.
- Friedman, J. (1997). On bias, variance, 0-1 loss and the curse of dimensionality, *J. Data Mining and Knowledge Discovery* **1**: 55–77.
- Friedman, J. (1999). Stochastic gradient boosting, *Technical report*, Stanford University.
- Friedman, J. (2001). Greedy function approximation: the gradient boosting machine, *Annals of Statistics* . to appear.
- Friedman, J., Baskett, F. and Shustek, L. (1975). An algorithm for finding nearest neighbors, *IEEE transactions on computers* **24**: 1000–1006.
- Friedman, J., Bentley, J. and Finkel, R. (1977). An algorithm for finding best matches in logarithmic expected time, *ACM Transactions on Mathematical Software* **3**: 209–226.

- Friedman, J. and Fisher, N. (1999). Bump hunting in high dimensional data, *Statistics and Computing* **9**: 123–143.
- Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion), *Annals of Statistics* **28**: 337–307.
- Friedman, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modelling (with discussion), *Technometrics* **31**: 3–39.
- Friedman, J. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**: 817–823.
- Friedman, J., Stuetzle, W. and Schroeder, A. (1984). Projection pursuit density estimation, *Journal of the American Statistical Association* **79**: 599–608.
- Friedman, J. and Tukey, J. (1974). A projection pursuit algorithm for exploratory data analysis, *IEEE trans. on computers, Ser. C* **23**: 881–889.
- Furnival, G. and Wilson, R. (1974). Regression by leaps and bounds, *Technometrics* **16**: 499–511.
- Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.* **85**: 398–409.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*, CRC Press, Boca Raton, FL.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.
- Gersho, A. and Gray, R. (1992). *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Boston, MA.
- Girosi, F., Jones, M. and Poggio, T. (1995). Regularization theory and neural network architectures, *Neural Computation* **7**: 219–269.
- Golub, G., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**: 215–224.
- Golub, G. and Van Loan, C. (1983). *Matrix Computations*, Johns Hopkins University Press, Baltimore.
- Gordon, A. (1999). *Classification (2nd edition)*, Chapman and Hall/CRC press, London.

- Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.
- Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*, Academic Press, New York.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
- Hand, D. (1981). *Discrimination and Classification*, Wiley, Chichester.
- Hart, P. (1968). The condensed nearest-neighbor rule, *IEEE Trans. Inform. Theory* **14**: 515–516.
- Hartigan, J. A. (1975). *Clustering Algorithms*, Wiley, New York.
- Hartigan, J. A. and Wong, M. A. (1979). [(Algorithm AS 136] A k -means clustering algorithm (AS R39: 81v30 p355-356), *Applied Statistics* **28**: 100–108.
- Hastie, T. (1984). Principal curves and surfaces, *Technical report*, Stanford University.
- Hastie, T., Botha, J. and Schnitzler, C. (1989). Regression with an ordered categorical response, *Statistics in Medicine* **43**: 884–889.
- Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *Annals of Statistics* **23**: 73–102.
- Hastie, T. and Herman, A. (1990). An analysis of gestational age, neonatal size and neonatal death using nonparametric logistic regression, *Journal of Clinical Epidemiology* **43**: 1179–90.
- Hastie, T. and Simard, P. (1998). Models and metrics for handwritten digit recognition, *Statistical Science* **13**: 54–65.
- Hastie, T. and Stuetzle, W. (1989). Principal curves, *Journal of the American Statistical Association* **84**(406): 502–516.
- Hastie, T. and Tibshirani, R. (1987). Nonparametric logistic and proportional odds regression, *Applied Statistics* **36**: 260–276.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall, London.
- Hastie, T. and Tibshirani, R. (1996a). Discriminant adaptive nearest-neighbor classification, *IEEE Pattern Recognition and Machine Intelligence* **18**: 607–616.

- Hastie, T. and Tibshirani, R. (1996b). Discriminant analysis by Gaussian mixtures, *J. Royal. Statist. Soc. B.* **58**: 155–176.
- Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling, *Annals of Statistics* **26**(2).
- Hastie, T., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring, *J. Amer. Statist. Assoc.* **89**: 1255–1270.
- Hastie, T., Tibshirani, R. and Buja, A. (1998). Flexible discriminant and mixture models, in J. Kay and M. Titterton (eds), *Statistics and Artificial Neural Networks*, Oxford University Press.
- Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions, *Statistics & Probability Letters* **4**: 53–56.
- Hebb, D. (1949). *The Organization of Behavior*, Wiley, New York.
- Hertz, J., Krogh, A. and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*, Addison Wesley, Redwood City, CA.
- Hinton, G. (1989). Connectionist learning procedures, *Artificial Intelligence* **40**: 185–234.
- Hoerl, A. E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**: 55–67.
- Huber, P. (1964). Robust estimation of a location parameter, *Annals of Math. Stat.* **53**: 73–101.
- Huber, P. (1985). Projection pursuit, *Annals of Statistics* **13**: 435–475.
- Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications, *Neural Networks* **13**: 411–430.
- Izenman, A. (1975). Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis* **5**: 248–264.
- Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts, *Neural computation* **3**: 79–87.
- Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*, Prentice-Hall.
- Jancey, R. (1966). Multidimensional group analysis, *Austral. J. Botany* **14**: 127–130.
- Jones, L. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Stat.* **20**: 608–613.

- Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation* **6**: 181–214.
- Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.
- Kearns, M. and Vazirani, U. (1994). *An Introduction to Computational Learning Theory*, MIT Press.
- Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism, *Biometrics* **46**: 1071–1085.
- Kohonen, T. (1989). *Self-Organization and Associative Memory (3rd edition)*, Springer-Verlag, Berlin.
- Kohonen, T. (1990). The self-organizing map, *Proc. of IEEE* **78**: 1464–1479.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, A. and Saarela, A. (2000). Self organization of a massive document collection, *IEEE Transactions on Neural Networks* **11**(3): 574–585. Special Issue on Neural Networks for Data Mining and Knowledge Discovery.
- Kressel, U. (1999). Pairwise classification and support vector machines, in B. Schölkopf, C. Burges and A. Smola (eds), *Advances in Kernel Methods - Support Vector Learning*, MIT Press, Cambridge, MA., pp. 255–268.
- Lawson, C. and Hansen, R. (1974). *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ.
- Le Cun, Y. (1989). Generalization and network design strategies, *Technical Report CRG-TR-89-4*, Dept. of Comp. Sci., Univ. of Toronto.
- Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1990). Handwritten digit recognition with a back-propagation network, in D. Touretzky (ed.), *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufman, Denver, CO.
- Le Cun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11): 2278–2324.
- Leblanc, M. and Tibshirani, R. (1996). Combining estimates in regression and classification, *J. Amer. Statist. Assoc.* **91**: 1641–1650.

- Lin, H., McCulloch, C., Turnbull, B., Slate, E. and Clark, L. (2000). A latent class mixed model for analyzing biomarker trajectories in longitudinal data with irregularly scheduled observations., *Statistics in Medicine* **19**: 1303–1318.
- Little, R. and Rubin, D. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
- Lloyd, S. (1957). Least squares quantization in PCM., *Technical report*, Bell Laboratories. Published in 1982 in *IEEE Trans. Inf. Theory* **28** 128-137.
- Loader, C. (1999). *Local Regression and Likelihood*, Springer–Verlag.
- Macnaughton Smith, P., Williams, W., Dale, M. and Mockett, L. (1965). Dissimilarity analysis: a new technique of hierarchical subdivision, *Nature* **202**: 1034–1035.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L.M. LeCam and J. Neyman, Univ. of California Press, pp. 281–297.
- Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty using occam's window., *J. Amer. Statist. Assoc.* **89**: 1535–46.
- Mardia, K., Kent, J. and Bibby, J. (1979). *Multivariate Analysis*, Academic Press.
- Massart, D., Plastria, F. and Kaufman, L. (1983). Non-hierarchical clustering with MASLOC, *The Journal of the Pattern Recognition Society* **16**: 507–516.
- McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* **5**: 115–133. pp 96-104; Reprinted in Andersen and Rosenfeld (1988).
- McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.
- Michie, D., Spiegelhalter, D. and Taylor, C. (eds) (1994). *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.
- Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association* pp. 415–434.

- Murray, W., Gill, P., and Wright, M. (1981). *Practical Optimization*, Academic Press.
- Myles, J. and Hand, D. (1990). The multiclass metric problem in nearest neighbor classification, *Pattern Recognition* **23**: 1291–1297.
- Neal, R. (1996). *Bayesian Learning for Neural Networks*, Springer-Verlag, New York.
- Neal, R. and Hinton, G. (1998). *A view of the EM algorithm that justifies incremental, sparse, and other variants; in Learning in Graphical Models, M. Jordan (ed.)*, Dordrecht: Kluwer Academic Publishers, Boston, MA., pp. 355–368.
- Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions, *Statistics & Probability Letters* **33**: 291–297.
- Parker, D. (1985). Learning logic, *Technical Report TR-87*, Cambridge MA: MIT Center for Research in Computational Economics and Management Science.
- Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization, in B. Schölkopf, C. J. C. Burges and A. J. Smola (eds), *Advances in Kernel Methods — Support Vector Learning*, MIT Press, Cambridge, MA., pp. 185–208.
- Quinlan, R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.
- Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*, Springer Verlag.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.
- Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.
- Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length, *Annals of Statistics* **11**: 416–431.
- Robbins, H. and Munro, S. (1951). A stochastic approximation method, *Ann. Math. Stat.* **22**: 400–407.
- Roosen, C. and Hastie, T. (1994). Automatic smoothing spline projection pursuit, *Journal of Computational and Graphical Statistics* **3**: 235–248.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological Review* **65**: 386–408.

- Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan, Washington, D.C.
- Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities, *South African Medical Journal* **64**: 430–436.
- Rumelhart, D., Hinton, G. and Williams, R. (1986). *Learning internal representations by error propagation in Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Rumelhart, D.E. and McClelland, J. L eds.)*, The MIT Press, Cambridge, MA., pp. 318–362.
- Schapire, R. (1990). The strength of weak learnability, *Machine Learning* **5**(2): 197–227.
- Schapire, R., Freund, Y., Bartlett, P. and Lee, W. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods, *Annals of Statistics* **26**(5): 1651–1686.
- Schapire, R. and Singer, Y. (1998). Improved boosting algorithms using confidence-rated predictions, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*.
- Schwartz, G. (1979). Estimating the dimension of a model, *Annals of Statistics* **6**: 461–464.
- Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York.
- Seber, G. (1984). *Multivariate Observations*, Wiley, New York.
- Shao, J. (1996). Bootstrap model selection, *J. Amer. Statist. Assoc.* **91**: 655–665.
- Short, R. and Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification, *IEEE Transactions on Information Theory* **27**: 622–627.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Silvey, S. (1975). *Statistical Inference*, Halsted.
- Simard, P., Le Cun, Y. and Denker, J. (1993). Efficient pattern recognition using a new transformation distance, *Advances in Neural Information Processing Systems*, Morgan Kaufman, San Mateo, CA, pp. 50–58.
- Spiegelhalter, D., Best, N., Gilks, W. and Inskip, H. (1996). Hepatitis B: a case study in MCMC methods, in W. Gilks, S. Richardson and D. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, Chapman and Hall, London.

- Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. radical prostatectomy treated patients, *Journal of Urology* **16**: 1076–1083.
- Stone, C., Hansen, M., Kooperberg, C. and Truong, Y. (1997). Polynomial splines and their tensor products (with discussion), *Annals of Statistics* **25**(4): 1371–1470.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. Roy. Statist. Soc.* **36**: 111–147.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, *J. Roy. Statist. Soc.* **39**: 44–7.
- Stone, M. and Brooks, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (Corr: V54 p906-907), *Journal of the Royal Statistical Society, Series B, Methodological* **52**: 237–269.
- Swayne, D., Cook, D. and Buja, A. (1991). Xgobi: Interactive dynamic graphics in the X window system with a link to S, *ASA Proceedings of Section on Statistical Graphics*, pp. 1–8.
- Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion), *J. Amer. Statist. Assoc.* **82**: 528–550.
- Tarpey, T. and Flury, B. (1996). Self-consistency: A fundamental concept in statistics, *Statistical Science* **11**: 229–243.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.* **58**: 267–288.
- Tibshirani, R. and Knight, K. (1999). Model search and inference by bootstrap “bumping”, *J. Comp. and Graph. Stat.* **8**: 671–686.
- Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic, *J. Royal. Statist. Soc. B.* (to appear) .
- Valiant, L. G. (1984). A theory of the learnable, *Communications of the ACM* **27**: 1134–1142.
- van der Merwe, A. and Zidek, J. (1980). Multivariate regression analysis and canonical variates, *The Canadian Journal of Statistics* **8**: 27–39.
- Vapnik, V. (1996). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

- Vidakovic, B. (1999). *Statistical Modeling by Wavelets*, Wiley, New York.
- Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data, *Proceedings of the International Conference on Approximation theory in Honour of George Lorenz*, Academic Press, Austin, Texas.
- Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.
- Wahba, G., Lin, Y. and Zhang, H. (2000). GACV for support vector machines, in A. Smola, P. Bartlett, B. Schölkopf and D. Schuurmans (eds), *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA., pp. 297–311.
- Weisberg, S. (1980). *Applied Linear Regression*, Wiley, New York.
- Werbos, P. (1974). *Beyond regression*, PhD thesis, Harvard University.
- Wickerhauser, M. (1994). *Adapted Wavelet Analysis from Theory to Software*, A.K. Peters Ltd, Natick, MA.
- Widrow, B. and Hoff, M. (1960). Adaptive switching circuits, Vol. 4, IRE WESCON Convention record. pp 96-104; Reprinted in Andersen and Rosenfeld (1988).
- Wold, H. (1975). Soft modelling by latent variables: The nonlinear iterative partial least squares (NIPALS) approach, *Perspectives in Probability and Statistics, In Honor of M. S. Bartlett*, pp. 117–144.
- Wolpert, D. (1992). Stacked generalization, *Neural Networks* **5**: 241–259.
- Yee, T. and Wild, C. (1996). Vector generalized additive models, *Journal of the Royal Statistical Society, Series B.* **58**: 481–493.
- Zhang, P. (1993). Model selection via multifold cross-validation, *Ann. Statist.* **21**: 299–311.