Statistical Society, Ser. B, 55, 39–52.
Gilks, W. R., and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Journal of the Royal Statistical Society*, Ser. C, 41, 337–348.
Ripley, B. D. (1987), *Stochastic Simulation*, New York: Wiley.
Ritter, C., and Tanner, M. A. (1992), "The Gibbs Stopper and the Griddy Gibbs Sampler," *Journal of the American Statistical Association*, 87, 861–868.
Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.
Rubin, D. B. (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, London: Oxford University Press, pp. 395–402.
Smith, A. F. M., and Gelfand, A. E. (1992), "Bayesian Statistics Without Tears," *American Statistician*, 46, 84–88.
Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Ser. B, 55, 3–23.
Spiegelhalter, D., Best, N., and Carlin, B. P. (1998), "Bayesian Deviance, the Effective Number of Parameters and the Comparison of Arbitrarily Complex Models," technical report, MRC Biostatistics Unit, Cambridge, U.K.
Spiegelhalter, D. J., Thomas, A., Best, N., and Gilks, W. R. (1995), "BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50," Medical Research Council, Biostatistics Unit, Cambridge, U.K.
Tanner, M. A. (1993), *Tools for Statistical Inference* (2nd ed.), New York: Springer-Verlag.
Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.
Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Ann. Statist.*, 22, 1701–1762.
Wakefield, J., Gelfand, A. E., and Smith, A. F. M. (1992), "Efficient Computation of Random Variates via the Ratio-of-Uniforms Method," *Statist. and Comput.*, 1, 129–133.
West, M. (1992), "Modeling With Mixtures," in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 503–524.

The Variable Selection Problem

Edward I. GEORGE

The problem of variable selection is one of the most pervasive model selection problems in statistical applications. Often referred to as the problem of subset selection, it arises when one wants to model the relationship between a variable of interest and a subset of potential explanatory variables or predictors, but there is uncertainty about which subset to use. This vignette reviews some of the key developments that have led to the wide variety of approaches for this problem.

1. INTRODUCTION

Suppose that $\mathbf{Y}$, a variable of interest, and $\mathbf{X}_1, \ldots, \mathbf{X}_p$, a set of potential explanatory variables or predictors, are vectors of n observations. The problem of variable selection, or subset selection as it is often called, arises when one wants to model the relationship between $\mathbf{Y}$ and a subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is thought to contain many redundant or irrelevant variables.

The variable selection problem is most familiar in the linear regression context, where attention is restricted to normal linear models. Letting γ index the subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ and letting q_γ be the size of the γth subset, the problem is to select and fit a model of the form

$$\mathbf{Y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \varepsilon, \tag{1}$$

where $\mathbf{X}_\gamma$ is an $n \times q_\gamma$ matrix whose columns correspond to the γth subset, $\boldsymbol{\beta}_\gamma$ is a $q_\gamma \times 1$ vector of regression coefficients, and $\varepsilon \sim N_n(0, \sigma^2 I)$. More generally, the variable selection problem is a special case of the model selection problem where each model under consideration corresponds to a distinct subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$. Typically, a single model class is simply applied to all possible subsets. For example, a wide variety of relationships can be considered with generalized linear models where $g(E(\mathbf{Y})) = \alpha + \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma$ for some link function g (see the vignettes by Christensen and McCulloch). Moving further away from the normal linear model, one might instead consider relating $\mathbf{Y}$ and subsets of $\mathbf{X}_1, \ldots, \mathbf{X}_p$ with nonparametric models such as CART or MARS.

The fundamental developments in variable selection seem to have occurred either directly in the context of the linear model (1) or in the context of general model selection frameworks. Historically, the focus began with the linear model in the 1960s, when the first wave of important developments occurred and computing was expensive. The focus on the linear model still continues, in part because its analytic tractability greatly facilitates insight, but also because many problems of interest can be posed as linear variable selection problems. For example, for the problem of nonparametric function estimation, $\mathbf{Y}$ represents the values of the unknown function, and $\mathbf{X}_1, \ldots, \mathbf{X}_p$ represent a linear basis, such as a wavelet basis or a spline basis. However, as advances in computing technology have allowed for the implementation of richer classes of models, treatments of the variable selection problem by general model selection approaches are becoming more prevalent.

One of the fascinating aspects of the variable selection problem has been the wide variety of methods that have

Edward I. George holds the Ed and Molly Smith Chair and is Professor of Statistics, Department of MSIS, University of Texas, Austin, TX 78712 (E-mail: *egeorge@mail.utexas.edu*). This work was supported by National Science Foundation grant DMS-98.03756 and Texas ARP grants 003658.452 and 003658.690.

been brought to bear on the problem. Because of space limitations, it is of course impossible to even mention them all, and so I focus on only a few to illustrate the general thrust of developments. An excellent and comprehensive treatment of variable selection methods prior to 1990 was provided by Miller (1990). As I discuss, many promising new approaches have appeared over the last decade.

2. GETTING A GRIP ON THE PROBLEM

A distinguishing feature of variable selection problems is their enormous size. Even with moderate values of p, computing characteristics for all 2^p models is prohibitively expensive, and some reduction of the model space is needed. Focusing on the linear model (1), early suggestions based such reductions on the residual sum of squares, which provided a partial ordering of the models. Taking advantage of the chain structure of subsets, branch and bound methods such as the algorithm of Furnival and Wilson (1974) were proposed to logically eliminate large numbers of models from consideration. When feasible, attention was often restricted to the "best subsets" of each size. Otherwise, reduction was obtained with variants of stepwise methods that sequentially add or delete variables based on greedy considerations (e.g., Efromyson 1966). Even with advances in computing technology, these methods continue to be the standard workhorses for reduction. Extensions beyond the linear model are straightforward; for example, in generalized linear models by substituting the deviance for the residual sum of squares.

Once attention was reduced to a manageable set of models, criteria were needed for selecting a subset model. The earliest developments of such selection criteria, again in the linear model context, were based on attempts to minimize the mean squared error of prediction. Different criteria corresponded to different assumptions about which predictor values to use, and whether they were fixed or random (see Hocking 1976; Thompson 1978 and the references therein). Perhaps the most familiar of those criteria is the Mallows $C_p = (\mathrm{RSS}_\gamma/\hat{\sigma}^2_{\mathrm{full}} + 2q_\gamma - n)$, where RSS_γ is the residual sum of squares for the γth model and $\hat{\sigma}^2_{\mathrm{full}}$ is the usual unbiased estimate of σ^2 based on the full model. Motivated as an unbiased estimate of predictive accuracy of the γth model, Mallows (1973) recommended using C_p plots to help gauge subset selection (see also Mallows 1995). Although Mallows specifically warned against using minimum C_p as a selection criterion (because of selection bias), minimum C_p continues to used as a criterion (and attributed to Mallows to boot!).

Two of the other most popular criteria, motivated from very different viewpoints, are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Letting $\hat{l}_\gamma$ denote the maximum log-likelihood of the γth model, AIC selects the model that maximizes $(\hat{l}_\gamma - q_\gamma)$, whereas BIC selects the model that maximizes $(\hat{l}_\gamma - (\log n)q_\gamma/2)$. Akaike (1973) motivated AIC from an information theoretic standpoint (see the vignette by Soofi) as the minimization of the Kullback–Leibler distance between the distributions of Y under the γth model and under the true model. To lend further support, an asymptotic equivalence of AIC and cross-validation was shown by Stone (1977). In contrast, Schwarz (1978) motivated BIC from a Bayesian standpoint, by showing that it was asymptotically equivalent (as $n \to \infty$) to selection based on Bayes factors. BIC was further justified from a coding theory viewpoint by Rissanen (1978).

Comparisons of the relative merits of AIC and BIC based on asymptotic consistency (as $n \to \infty$) have flourished in the literature. As it turns out, BIC is consistent when the true model is fixed (Haughton 1998), whereas AIC is consistent if the dimensionality of the true model increases with n (at an appropriate rate) (Shibata 1981). Stone (1979) provided an illuminating discussion of these two viewpoints.

For the linear model (1), many of the popular selection criteria are special cases of a penalized sum of squares criterion, providing a unified framework for comparisons. Assuming σ^2 known to avoid complications, this general criterion selects the subset model that minimizes

$$(\mathrm{RSS}_\gamma/\hat{\sigma}^2 + Fq_\gamma), \tag{2}$$

where F is a preset "dimensionality penalty." Intuitively, (2) penalizes $\mathrm{RSS}_\gamma/\hat{\sigma}^2$ by F times q_γ, the dimension of the γth model. AIC and minimum C_p are essentially equivalent, corresponding to $F = 2$, and BIC is obtained by setting $F = \log n$. By imposing a smaller penalty, AIC and minimum C_p will select larger models than BIC (unless n is very small).

3. TAKING SELECTION INTO ACCOUNT

Further insight into the choice of F is obtained when all of the predictors are orthogonal, in which case (2) simply selects all of those predictors with t-statistics t for which $t^2 > F$. When $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are in fact all unrelated to $\mathbf{Y}$ (i.e., the full model regression coefficients are all 0), AIC and minimum C_p are clearly too liberal and tend to include a large proportion of irrelevant variables. A natural conservative choice for F, namely $F = 2\log p$, is suggested by the fact that under this null model, the expected value of the largest squared t-statistic is approximately $2\log p$ when p is large. This choice is the risk inflation criterion (RIC) proposed by Foster and George (1994) and the universal threshold for wavelets proposed by Donoho and Johnstone (1994). Both of these articles motivate $F = 2\log p$ as yielding the smallest possible maximum inflation in predictive risk due to selection (as $p \to \infty$), a minimax decision theory standpoint. Motivated by similar considerations, Tibshirani and Knight (1999) recently proposed the covariance inflation criterion (CIC), a nonparametric method of selection based on adjusting the bias of in-sample performance estimates. Yet another promising adjustment based on a generalized degrees of freedom concept was proposed by Ye (1998).

Many other interesting criteria corresponding to different choices of F in (2) have been proposed in the literature (see, e.g., Hurvitz and Tsai 1989, 1998; Rao and Wu 1989; Shao 1997; Wei 1992; Zheng and Loh 1997 and the references therein). One of the drawbacks of using a fixed choice of F, is that models of a particular size are favored; small F favors large models, and large F favors small models. Adaptive choices of F to mitigate this problem have been

recommended by Benjamini and Hochberg (1995), Clyde and George (1999, 2000), George and Foster (2000), and Johnstone and Silverman (1998).

An alternative to explicit criteria of the form (2), is selection based on predictive error estimates obtained by intensive computing methods such as the bootstrap (e.g., Efron 1983; Gong 1986) and cross-validation (e.g., Shao 1993; Zhang 1993). An interesting variant of these is the little bootstrap (Brieman 1992), which estimates the predictive error of selected models by mimicking replicate data comparison. The little bootstrap compares favorably to selection based on minimum C_p or the conditional bootstrap, whose performances are seriously denigrated by selection bias.

Another drawback of traditional subset selection methods, which is beginning to receive more attention, is their instability relative to small changes in the data. Two novel alternatives that mitigate some of this instability for linear models are the nonnegative garrotte (Brieman 1995) and the lasso (Tibshirani 1996). Both of these procedures replace the full model least squares criterion by constrained optimization criteria. As the constraint is tightened, estimates are zeroed out, and a subset model is identified and estimated.

4. BAYESIAN METHODS EMERGE

The fully Bayesian approach to variable selection is as follows (George 1999). For a given set of models $M_1, \ldots, M_{2^p}$, where M_γ corresponds to the γth subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$, one puts priors $\pi(\beta_\gamma | M_\gamma)$ on the parameters of each M_γ and a prior on the set of models $\pi(M_1), \ldots, \pi(M_{2^p})$. Selection is then based on the posterior model probabilities $\pi(M_\gamma | Y)$, which are obtained in principle by Bayes's theorem.

Although this Bayesian approach appears to provide a comprehensive solution to the variable selection problem, the difficulties of prior specification and posterior computation are formidable when the set of models is large. Even when p is small and subjective considerations are not out of the question (Garthwaite and Dickey 1995), prior specification requires considerable effort. Instead, many of the Bayesian proposals have focused on semiautomatic methods that attempt to minimize prior dependence. Indeed, this is part of the appeal of BIC, which avoids prior specification altogether, and its properties continue to be investigated and justified (Kass and Wasserman 1995; Pauler 1998; Raftery 1996). Other examples of Bayesian treatments that avoid the prior selection difficulties in variable selection include the early proposal of Lindley (1968) to use uniform priors and a cost function for selection, the default Bayes factor criteria of Berger and Pericchi (1996a,b) and O'Hagan (1995), and the predictive criteria of Geisser and Eddy (1979), Laud and Ibrahim (1995), and San Martini and Spezzaferri (1984).

In contrast to the development of Bayesian approaches that avoid the difficulties of prior specification, the advent of Markov chain Monte Carlo (MCMC) (see the vignette by Cappe and Robert) has focused attention on Bayesian variable selection with fully specified proper parameter priors. Bypassing the difficulties of computing the entire posterior, MCMC algorithms can instead be used to stochastically search for the high-posterior probability models. The idea is that by simulating a Markov chain, which is converging to the posterior distribution, the high-probability models should tend to appear more often, and hence sooner. The resulting implementations are stepwise algorithms that are stochastically guided by the posterior, rather than by the greedy considerations of conventional stepwise methods. Such a Bayesian package is complete; it offers posterior probability as a selection criteria, associated MCMC algorithms for search, and Bayes estimates for the selected model.

The last decade has seen an explosion of research on this Bayesian variable selection approach. These developments have included proposals for new prior specifications that induce increased posterior probability on the more promising models, for new MCMC implementations that are more versatile and offer improved performance, and for extensions to a wide variety of model classes. Another closely related development in this context has been the emergence of model averaging as an alternative to variable selection. Under the Bayesian variable selection formulation, the posterior mean is an adaptive convex combination of all the individual model estimates (i.e., a model average). Although model averaging almost always improves on variable selection in terms of prediction, its drawback is that it does not lead to a reduced set of variables. Some, but by no means all, of the key developments of these Bayesian approaches to variable selection and model averaging have been discussed by Clyde (1999), Clyde, Parmigiani, and Vidakovic (1998), Draper (1995), George and McCulloch (1993, 1997), Green (1995), and Hoeting, Madigan, Raftery, and Volinsky (1999).

5. WHAT IS NEXT

Today, variable selection procedures are an integral part of virtually all widely used statistics packages, and their use will only increase as the information revolution brings us larger datasets with more and more variables. The demand for variable selection will be strong, and it will continue to be a basic strategy for data analysis.

Although numerous variable selection methods have been proposed, plenty of work still remains to be done. To begin with, many of the recommended procedures have been given only a narrow theoretical motivation, and their operational properties need more systematic investigation before they can be used with confidence. For example, small-sample justification is needed in addition to asymptotic considerations, and frequentist justification is needed for Bayesian procedures. Although there has been clear progress on the problems of selection bias, clear solutions are still needed, especially for the problems of inference after selection (see Zhang 1992). Another intriguing avenue for research is variable selection using multiple model classes (see Donoho and Johnstone 1995). New problems will also appear as demand increases for data mining of massive datasets. For example, considerations of scalability and computational efficiency will become paramount in such a context. I suppose that all of this is good news, but there is also danger lurking ahead.

With the availability of so many variable selection procedures and so many different justifications, it has becomes increasingly easy to be misled and to mislead. Faced with too many choices and too little guidance, practitioners continue to turn to the old standards such as stepwise selection based on AIC or minimum C_p, followed by a report of the conventional estimates and inferences. The justification of asymptotic consistency will not help the naive user who should be more concerned with selection bias and procedure instability. Eventually, the responsibility for the poor performance of such procedures will fall on the statistical profession, and consumers will turn elsewhere for guidance (e.g., Dash and Liu 1997). Our enthusiasm for the development of promising new procedures must be carefully tempered with cautionary warnings of their potential pitfalls.

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Berger, J. O., and Pericchi, L. R. (1996a), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

——— (1996b), "The Intrinsic Bayes Factor for Linear Models," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 25–44.

Brieman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

——— (1995), "Better Subset Selection Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Clyde, M. (1999), "Bayesian Model Averaging and Model Search Strategies," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press.

Clyde, M., and George, E. I. (1999), "Empirical Bayes Estimation in Wavelet Nonparametric Regression," in *Bayesian Inference in Wavelet-Based Models*, eds. P. Muller and B. Vidakovic, New York: Springer-Verlag, pp. 309–322.

——— (2000), "Flexible Empirical Bayes Estimation for Wavelets," *Journal of the Royal Statistical Society*, Ser. B, 681–698.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), "Multiple Shrinkage and Subset Selection in Wavelets," *Biometrika*, 85, 391–402.

Dash, M., and Liu, H. (1997), "Feature Selection for Classification," *Intelligent Data Analysis*, 1, 131–156.

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–256.

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the Royal Statistical Society*, Ser. B, 90, 1200–1224.

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 45–97.

Efron, B. (1983), "Estimating the Error Rate of a Predictive Rule: Improvement Over Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

Efroymson, M. A. (1960), "Multiple Regression Analysis," in *Mathematical Methods for Digital Computers*, eds. A. Ralston and H. S. Wilf, New York: Wiley, pp. 191–203.

Foster, D. P., and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Furnival, G. M., and Wilson, R. W. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

Garthwaite, P. H., and Dickey, J. M. (1996), "Quantifying and Using Expert Opinion for Variable-Selection Problems in Regression" (with discussion), *Chemometrics and Intelligent Laboratory Systems*, 35, 1–34.

Geisser, S., and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

George, E. I. (1999), "Bayesian Model Selection," in *Encyclopedia of Statistical Sciences, Update Vol. 3*, eds. S. Kotz, C. Read, and D. Banks, New York: Wiley, pp. 39–46.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87 (forthcoming).

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7(2), 339–373.

Gong, G. (1986), "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," *Journal of the American Statistical Association*, 393, 108–113.

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Haughton, D. (1988), "On the Choice of a Model to Fit Data From an Exponential Family," *The Annals of Statistics*, 16, 342–355.

Hocking (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–417.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

——— (1998), "A Cross-Validatory AIC for Hard Wavelet Thresholding in Spatially Adaptive Function Estimation," *Biometrika*, 85, 701–710.

Johnstone, I. M., and Silverman, B. W. (1998), "Empirical Bayes Approaches to Mixture Problems and Wavelet Regression," technical report, University of Bristol.

Kass, R. E., and Wasserman, L. (1995), "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90, 928–934.

Laud, P. W., and Ibrahim, J. G. (1995), "Predictive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 57, 247–262.

Lindley, D. V. (1968), "The Choice of Variables in Multiple Regression" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 30, 31–66.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–676.

——— (1995), "More Comments on C_p," *Technometrics*, 37, 362–372.

Miller, A. (1990), *Subset Selection in Regression*, London: Chapman and Hall.

O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.

Pauler, D. (1998), "The Schwarz Criterion and Related Methods for the Normal Linear Model," *Biometrika*, 85, 13–27.

Raftery, A. E. (1996), "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models," *Biometrika*, 83, 251–266.

Rao, C. R., and Wu, Y. (1989), "A Strongly Consistent Procedure for Model Selection in a Regression Problem," *Biometrika*, 76, 369–374.

Rissanen, J. (1978), "Modeling by Shortest Data Description," *Automatica*, 14, 465–471.

San Martini, A., and Spezzaferri, F. (1984), "A Predictive Model Selection Criterion," *Journal of the Royal Statistical Society*, Ser. B, 46, 296–303.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica*, 7(2), 221–264.

——— (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

——— (1979), "Comments on Model Selection Criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society*, Ser. B, 41, 276–278.

Thompson, M. L. (1978), "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation," *International Statistical Review*, 46, 1–19.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tibshirani, R., and Knight, K. (1999), "The Covariance Inflation Criterion for Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 61, 529–546.

Wei, C. Z. (1992), "On Predictive Least Squares Principles," *The Annals of Statistics*, 29, 1–42.

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120-131.

Zhang, P. (1992), "Inference After Variable Selection in Linear Regression Models," *Biometrika*, 79, 741–746.

——— (1993), "Model Selection via Multifold Cross-Validation," *The Annals of Statistics*, 21, 299–313.

Zheng, X., and Loh, W. Y. (1997), "A Consistent Variable Selection Criterion for Linear Models for Linear Models With High-Dimensional Covariates," *Statistica Sinica*, 7(2), 311–325.

Robust Nonparametric Methods

Thomas P. HETTMANSPERGER, Joseph W. MCKEAN, and Simon J. SHEATHER

1. A SHORT HISTORY OF NONPARAMETRICS FOR LOCATION PROBLEMS

The terms "nonparametric statistics" and "distribution-free methods" have historically referred to a collection of statistical tests whose null distributions do not depend on the underlying distribution of the data. The area has been expanded to include estimates and confidence intervals derived from the tests and includes asymptotically distribution-free tests for complex models as well. It is now generally recognized that these statistical methods are highly efficient over large sets of possible models and are robust as well. This was not always the case. The earliest work was by Hotelling and Pabst in 1936 (references to works in this paragraph can be found in the books cited in the next paragraph) on rank correlation and by Friedman in 1937 on rank tests in a two-way design, followed in 1945 with the introduction of the signed rank and rank sum tests by Wilcoxon. In 1947, Mann and Whitney extended the ideas of Wilcoxon. All of these rank tests were considered quick and dirty but not competitive with the presumably more efficient t-tests. This perspective was about to change. In 1948, Pitman developed efficiency concepts in a set of lecture notes that, although never published, were widely disseminated. During the 1950s and 1960s, Lehmann, working with Hodges and with his students at Berkeley, showed that rank tests are surprisingly efficient and robust. In fact, the Wilcoxon rank tests are essentially efficient at the normal model and can be much more efficient than least squares methods when the underlying distribution of the data has heavy tails. In the early 1960s, Hodges and Lehmann derived estimates and confidence intervals from rank test statistics and also introduced aligned rank tests for use in the regression model. In 1967, Hájek and Šidák published their seminal book on the theory of rank tests.

It is impossible to mention all of the contributions to this early development. Hence we cite several key monographs that contain many references to this work and that extended these methods into various settings. Puri and Sen (1971) published a monograph on nonparametric methods in multivariate models. This work extended and applied another major work on asymptotic theory by Chernoff and Savage (1958). Lehmann's (1975) book combined applied nonparametric methods with an extensive appendix on theory. The decade closed with the publication of Randles and Wolfe's (1979) introduction to the theory of nonparametric statistics, which provided a set of tools for the development of nonparametric methods.

2. RANK-BASED PROCEDURES FOR LINEAR MODELS

The nonparametric procedures (testing and estimation) for simple location problems offer the user highly efficient and robust methods and form an attractive alternative to traditional least squares (LS) procedures. LS procedures, however, generalize easily to any linear model and to most nonlinear models. The LS procedures are not model dependent. In contrast, there are very few classical nonparametric procedures for designs other than the simple location designs, and, furthermore, these procedures vary with the problem. For instance, the ranking procedure for the Kruskal–Wallis test of treatment effect in a one-way layout is much different than the ranking procedure for the Friedman test of treatment effect in a two-way design. Further, the efficiencies of these two procedures differ widely, although both are based on linear rankings.

The generality of LS procedures can readily be seen in terms of its simple geometry. For example, suppose that a vector of responses $\mathbf{Y}$ follows a linear model of the form $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{e}$, where $\mathbf{e}$ is a vector of random errors, $\boldsymbol{\mu} \in \Omega$, and Ω is a subspace of R^n. The LS estimate of $\boldsymbol{\mu}$ is the vector $\hat{\mathbf{Y}}_{\mathrm{LS}}$ that lies "closest" to $\mathbf{Y}$ when distance is Euclidean; that is, $\|\mathbf{Y} - \hat{\mathbf{Y}}_{\mathrm{LS}}\|_{\mathrm{LS}} = \min \|\mathbf{Y} - \boldsymbol{\mu}\|_{\mathrm{LS}}$ over all $\boldsymbol{\mu} \in \Omega$, where $\|\cdot\|_{\mathrm{LS}}$ denotes the Euclidean norm. The F test of $H_0 : \boldsymbol{\mu} \in \omega$ versus $H_A : \boldsymbol{\mu} \in \Omega \cap \omega^{\perp}$ is based on the standardized difference in squared distances between $\mathbf{Y}$ and each of the subspaces $\omega \subset \Omega$. These few sentences on

Thomas P. Hettmansperger is Professor of Statistics, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *tph@stat.psu.edu*). Joseph W. McKean is Professor of Statistics, Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008 (E-mail: *joe@stat.wmich.edu*). Simon J. Sheather is Professor, Australian Graduate School of Management, University of New South Wales, Sydney, NSW 2052, Australia (E-mail: *simonsh@agsm.unsw.edu.au*).

Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes