The Variable Selection Problem

Edward I. George *
University of Texas at Austin

September 2000

Abstract

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 which have led to the wide variety of approaches for this problem.

1 Introduction

Suppose Y a variable of interest, and $X_1, \ldots, X_p$ 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 Y and a subset of $X_1, \ldots, X_p$, but there is uncertainty about which subset to use. Such a situation is particularly of interest when p is large and $X_1, \ldots, X_p$ 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.

*Edward I. George holds the Ed and Molly Smith Chair and is Professor of Statistics, Dept. of MSIS, CBA 5.202, University of Texas, Austin, TX 78712-1175, egeorge@mail.utexas.edu. This work was supported by NSF grant DMS-98.03756 and Texas ARP grants 003658.452 and 003658.690.

Letting γ index the subsets of $X_1, \ldots, X_p$ and letting q_γ be the size of the γth subset, the problem is to select and fit a model of the form

$$Y = X_\gamma \beta_\gamma + \epsilon \tag{1}$$

where X_γ is an n x q_γ matrix whose columns correspond to the γth subset, β_γ is a q_γ x 1 vector of regression coefficients and $\epsilon \sim N_n(0, \sigma^2 I)$. 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 $X_1, \ldots, X_p$. 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 $g(E(Y)) = \alpha + X_\gamma \beta_\gamma$ 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 Y and subsets of $X_1, \ldots, X_p$ 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, Y represents the values of the unknown function, and $X_1, \ldots, X_p$ 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 is the wide variety of methods that have been brought to bear on the problem. Because of space limitations, it will of course be impossible to even mention them all, and so I have only focused on a few to illustrate the general thrusts of developments. An excellent and comprehensive treatment of variable selection methods prior to 1990 can be found in Miller (1990). As we will see, 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 today's 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 to select a subset model. The earliest developments of such selection criteria, again in the the linear model context, were based on attempts to minimize the mean square 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) and Thompson (1978) and the references therein. Perhaps the most familiar of those criteria is Mallows $C_p = (RSS_\gamma/\hat{\sigma}^2_{FULL} + 2q_\gamma - n)$, where RSS_γ is the residual sum-of squares for the γth model and $\hat{\sigma}^2_{FULL}$ 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 the use of C_p plots to help gauge subset selection, see also Mallows (1995). Although he specifically warned against 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 points of view, are AIC (for Akaike Information Criterion) and BIC (for Bayesian Information Criterion). Letting $\hat{\ell}_\gamma$ denote the maximum log likelihood of the γth model, AIC selects the model which maximizes $(\hat{\ell}_\gamma - q_\gamma)$, whereas BIC selects the model which maximizes

$(\hat{\ell}_\gamma - (\log n)q_\gamma/2)$. Akaike motivated AIC from an information theoretic point of view (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 point of view, 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 point of view 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) provides an illuminating discussion of these two points of view.

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

$$(RSS_\gamma/\hat{\sigma}^2 + F\, q_\gamma) \tag{2}$$

where F is a preset "dimensionality penalty". Intuitively, (2) penalizes $RSS_\gamma/\hat{\sigma}^2$ 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 above is obtained when all the predictors are orthogonal, in which case (2) simply selects all those predictor with t-statistics t for which $t^2 > F$. When $X_1, \ldots, X_p$ are in fact all unrelated to Y (i.e. the full model regression coefficients are all zero), 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 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 suggests the choice $F = 2\log p$, which corresponds to 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 papers 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 point of view. Motivated by similar considerations, Tibshirani mad 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 for example Hurvitz and Tsai (1989, 1998), Rao and Wu (1989), Wei (1992), Shao (1997), 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 which 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 $X_1, \ldots, X_p$, 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 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 semi-automatic methods which 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), Raftery (1996) and Pauler (1998). Other examples of Bayesian treatments which 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 O'Hagan (1995) and Berger and Pericchi (1996ab), and the predictive criteria of Geisser and Eddy (1979), San Martini and Spezzaferri (1984) and Laud and Ibrahim (1995).

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 which 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 can be found in George and McCulloch (1993), Draper (1995), Green (1995), George and McCulloch (1997), Clyde, Parmigiani and Vidakovic (1998), Clyde (1999), Hoeting, Madigan, Raftery and Volinsky (1999) and the references therein.

5 What's next

Today, variable selection procedures are an integral part of virtually all widely used statistics package, and their use will only increase as the information revolution brings us larger data sets 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 a wide variety of variable selection methods have been proposed, there is still plenty of work to be done. To begin with, many of the recommended procedures have been given a 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. While 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 data sets. For example,

considerations of scalability and computational efficiency will become paramount in such a context. I suppose 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 the instability of the procedures. 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 by 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, pp. 267-81. Budapest: Akademia Kiado.

Benjamini, Y. & Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Ser. B*, 57, 289-300.

Berger, J.O. and Pericchi, L.R. (1996a) The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, 91, 109-122.

Berger, J.O. and Pericchi, L.R. (1996b) The intrinsic Bayes factor for linear models. *Bayesian Statistics 5* J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Oxford: University Press, 25-44.

Brieman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Amer. Statist. Assoc.*, 87, 738-754.

Brieman, L. (1995) Better subset selection using the nonnegative garrote. *Technometrics*, 37, 373-384.

Clyde, M. (1999) Bayesian model averaging and model search strategies. *Bayesian Statistics 6* J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.), Oxford: University Press.

Clyde, M. and George, E.I. (1999) Empirical Bayes estimation in wavelet nonparametric regression. *Bayesian Inference in Wavelet Based Models* (eds. P. Muller and B. Vidakovic). Springer-Verlag., 309-322.

Clyde, M. and George, E.I. (2000) Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistics Society, Series B*, (forthcoming).

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* 131-156.

Donoho, D.L., and Johnstone, I.M. (1995), Adapting to unknown smoothness via wavelet shrinkage. *J. Royal Statist. Soc., Ser. B*, 90, 1200-1224.

Donoho, D.L., and Johnstone, I.M. (1994), Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-256.

Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *J. Royal Statist. Soc., Ser. B*, 57, 45-97.

Efron, B. (1983) Estimating the error rate of a predictive rule: Improvement over cross-validation. *J. Amer. Statist. Assoc.*, 78, 316-331.

Efroymson, M.A. (1960), Multiple regression analysis. *Mathematical Methods for Digital Computers*, (eds A. Ralston and H.S. Wilf). Wiley, New York 191-203.

Foster, D.P., and George, E.I. (1994), The risk inflation criterion for multiple regression. *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. *J. Amer. Statist. Assoc.* 74, 153-160.

George, E.I. (1999) Bayesian model selection. *Encyclopedia of Statistical Sciences, Update Volume 3*, (eds. S. Kotz, C. Read and D. Banks), Wiley, N.Y., 39-46

George, E.I. and Foster, D.P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, (forthcoming).

George, E.I., and McCulloch, R.E. (1993), Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, 88, 881-889.

George, E.I., and McCulloch, R.E. (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. *J. Amer. Statist. Assoc.*, 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. *Ann. Statist.* 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.

Hurvich, C.M. and Tsai, C.L. (1998) A crossvalidatory *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. Tech 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. *J. Amer. Statist. Assoc.* 90, 928-934.

Laud, P.W. and Ibrahim, J.G. (1995) Predictive model selection. *Journal of the Royal Statistics Society, Series B*, 57 247-262.

Lindley, D.V. (1968), The choice of variables in multiple regression (with discussion), *Journal of the Royal Statistics Society, Series B*, 30, 31-66.

Mallows, C.L. (1973), Some comments on C_p. *Technometrics*, 15, 661-676.

Mallows, C.L. (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 Statistics Society, Series 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 *J. Royal Statist. Soc., Ser. B*, 46 296-303.

Schwarz, G. (1978), Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Shao, J. (1997) An asymptotic theory for linear model selection. *Statistica Sinica*, 7, 2, 221-264.

Shao, J. (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 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. *J. Royal Statist. Soc., Ser. B*, 39, 44-47.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *J. Royal Statist. Soc., 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. *J.R. Statist. Soc. B* 58 267-288.

Tibshirani, R. and Knight, K. (1999) The covariance inflation criterion for model selection. *J. Royal Statist. Soc., Ser. B*,

Wei, C.Z. (1992) On predictive least squares principles. *Ann. Statist.* 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, Ping (1992) Inference after variable selection in linear regression models. *Biometrika*, 79, 741-746.

Zhang, Ping (1993) Model selection via multifold cross-validation. *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.