University of Pennsylvania
ScholarlyCommons

Statistics Papers

Wharton Faculty Research

12-2000

Calibration and Empirical Bayes Variable Selection

Edward I. George

Dean P. Foster
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/statistics_papers

Part of the Statistics and Probability Commons

Recommended Citation

George, E. I., & Foster, D. P. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika, 87* (4), 731-747. http://dx.doi.org/10.1093/biomet/87.4.731

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/statistics_papers/567
For more information, please contact repository@pobox.upenn.edu.

Calibration and Empirical Bayes Variable Selection

Abstract

For the problem of variable selection for the normal linear model, selection criteria such as AIC, C_p, BIC and RIC have fixed dimensionality penalties. Such criteria are shown to correspond to selection of maximum posterior models under implicit hyperparameter choices for a particular hierarchical Bayes formulation. Based on this calibration, we propose empirical Bayes selection criteria that use hyperparameter estimates instead of fixed choices. For obtaining these estimates, both marginal and conditional maximum likelihood methods are considered. As opposed to traditional fixed penalty criteria, these empirical Bayes criteria have dimensionality penalties that depend on the data. Their performance is seen to approximate adaptively the performance of the best fixed-penalty criterion across a variety of orthogonal and nonorthogonal set-ups, including wavelet regression. Empirical Bayes shrinkage estimators of the selected coefficients are also proposed.

Keywords

AIC, BIC, conditional likelihood, Cp, hierarchical model, marginal likelihood, model selection, RIC, Risk, Selection bias, shrinkage estimation, wavelets

Disciplines

Statistics and Probability

This journal article is available at ScholarlyCommons: https://repository.upenn.edu/statistics_papers/567

Calibration and Empirical Bayes Variable Selection

By EDWARD I. GEORGE
Department of MSIS, The University of Texas at Austin, Austin, TX 78712-1175, U.S.A.
egeorge@mail.utexas.edu

and DEAN P. FOSTER
Department of Statistics, The Wharton School of The University of Pennsylvania, Philadelphia, PA 19104-6302, U.S.A.
foster@diskworld.wharton.upenn.edu

SUMMARY

For the problem of variable selection for the normal linear model, selection criteria such as AIC, C_p, BIC and RIC have fixed dimensionality penalties. Such criteria are shown to correspond to selection of maximum posterior models under implicit hyperparameter choices for a particular hierarchical Bayes formulation. Based on this calibration, we propose empirical Bayes selection criteria that use hyperparameter estimates instead of fixed choices. For obtaining these estimates, both marginal and conditional maximum likelihood methods are considered. As opposed to traditional fixed penalty criteria, these empirical Bayes criteria have dimensionality penalties that depend on the data. Their performance is seen to approximate adaptively the performance of the best fixed penalty criterion across a variety of orthogonal and nonorthogonal set-ups, including wavelet regression. Empirical Bayes shrinkage estimators of the selected coefficients are also proposed.

Some keywords: AIC; BIC; Conditional likelihood; C_p; Hierarchical models; Marginal likelihood; Model selection; RIC; Risk; Selection bias; Shrinkage estimation; Wavelets.

1 Introduction

We consider the following canonical variable selection problem. Suppose we have n observations on a dependent variable Y and p independent variables $X = (x_1, \ldots, x_p)$, for which the familiar normal linear model holds:

$$Y = X\beta + \epsilon, \tag{1}$$

where $\epsilon \sim N_n(0, \sigma^2 I)$ and $\beta = (\beta_1, \ldots, \beta_p)'$. Suppose also that only an unknown subset of the β_j coefficients are nonzero. The problem of variable selection, or subset selection as it is often called, is to identify this unknown subset. The special case of this problem where $X = I$ occurs naturally in the context of wavelet regression; see Donoho & Johnstone (1994).

A common strategy for this variable selection problem has been to select the model that maximises a penalised sum of squares criterion. More precisely, if $\gamma = 1, \ldots, 2^p$ indexes all the subsets of $x_1, \ldots, x_p$, let

$$SS_\gamma = \hat{\beta}_\gamma' X_\gamma' X_\gamma \hat{\beta}_\gamma, \tag{2}$$

where X_γ is the $n \times q_\gamma$ matrix whose columns are the q_γ variables in the γth subset, and $\hat{\beta}_\gamma \equiv (X_\gamma' X_\gamma)^{-1} X_\gamma' Y$ is the least squares estimate of the vector of coefficients β_γ for this subset. Here SS_γ is just the familiar regression sum of squares for the γth model. The typical penalised sum of squares criterion entails picking the γth model that maximises

$$SS_\gamma / \hat{\sigma}^2 - F\, q_\gamma, \tag{3}$$

or equivalently minimises $SSE_\gamma / \hat{\sigma}^2 + F\, q_\gamma$ for $SSE_\gamma = Y'Y - SS_\gamma$. In (3), F is a fixed constant and $\hat{\sigma}^2$ is an estimator of σ^2. F can be interpreted as a 'dimensionality penalty' in the sense that (3) penalises $SS_\gamma / \hat{\sigma}^2$ by F times q_γ, the dimension of the γth model. Common choices for $\hat{\sigma}^2$ include $(Y'Y - SS_{\text{FULL}})/(n - p)$, the traditional unbiased estimator based on the full model, and $SSE_\gamma/(n - q_\gamma)$. In the wavelet regression context where $p = n$, Donoho et al. (1995) recommend the median absolute deviation estimator.

A number of popular criteria correspond to using (3) with different choices of F. Perhaps the most widely used are $F = 2$ which yields C_p (Mallows, 1973) and, approximately, AIC (Akaike, 1973), and $F = \log n$ which yields BIC (Schwarz, 1978). The motivation for these choices are varied; C_p is motivated as an unbiased estimate of predictive risk, AIC by an expected information distance, and BIC as an asymptotic Bayes factor. A wide variety of related choices for F have been proposed by others.

A variant of the above criterion, which plays an important role in this paper, is that with $F = 2 \log p$. This choice was proposed by Foster & George (1994) where it was called the Risk Inflation Criterion (RIC) because it asymptotically minimises the maximum predictive risk inflation due to selection when X is orthogonal. This choice and its minimax property were also discovered independently by Donoho & Johnstone (1994) in the wavelet regression context, where they refer to it as the universal hard thresholding rule.

Although initially motivated by minimaxity considerations, RIC can also be motivated by its relationship to the expected size of the largest t-statistic under the null model. More precisely, when X is orthogonal and $\beta_1 = \cdots = \beta_p = 0$, it can be shown that the expected size of the maximum squared t-statistic is approximately $2 \log p$ for large p. Since adding a variable to the model increases $SS_\gamma/\hat{\sigma}^2$ by the square of its t-statistic, when $\hat{\sigma}^2 \equiv s^2$, RIC essentially selects only those variables whose squared t-statistics exceed the expectation of the maximum under the null model. Although this is an intuitively reasonable adjustment for selection, this motivation also suggests that RIC is conservative since, under the null model, we would expect the size of the smaller squared t-statistics to be less than $2 \log p$. Indeed, under the null model with orthogonal predictors, it can be shown that the expected size of the qth largest squared t-statistic is approximately $2 \log(p/q)$ for large p. Based on this, one might consider the modified RIC criterion

$$SS_\gamma/\hat{\sigma}^2 - \sum_{q=1}^{q_\gamma} 2 \log(p/q), \tag{4}$$

that compares the ordered t-statistics to their expected size under the null model. Note that both RIC and (4) exploit aspects of the expected ensemble of t-statistics. However, in contrast to RIC and criteria of the form (3) which use a fixed penalty F, (4) uses the penalty $2 \log(p/q)$ which decreases as q increases. Benjamini & Hochberg (1995) have proposed similar modifications of the $2 \log p$ threshold for multiple comparisons problems. More recently, information-theoretic motivations for such modifications have been explored by Foster & Stine (1999) and Hansen & Yu (1999).

Alternatives to criteria of the form (3) and (4) are provided by a Bayesian approach. The basic idea is to put a prior distribution on β_γ and γ, and then to use the posterior distribution $p(\gamma \mid Y)$ to identify models. An appealing aspect of such an approach is that it automatically takes ensemble information into account. A key element, of course, is the choice of the prior distribution $p(\beta_\gamma, \gamma)$. Although subjective considerations can occasionally be used, see Dickey & Garthwaite (1996), the complexity of required inputs often necessitate 'default' choices that, one hopes, lead to posteriors that put high probability on the 'most promising' models. Many variants of the Bayesian approach to variable selection have been proposed in the literature; see Mitchell & Beauchamp (1988) and George & McCulloch (1993, 1995, 1997) for further references. When the goal is exclusively prediction, the Bayesian approach also provides a useful alternative to variable selection called Bayesian model averaging; see Clyde, Desimone & Parmigiani (1996) and Raftery, Madigan & Hoeting (1997).

Recent advances in Markov chain Monte Carlo methods for posterior computation have led to increased interest in all of these methods.

In this paper, we propose and develop empirical Bayes alternatives to criteria of the form (3). In § 2, we show how a popular hierarchical Bayes formulation for variable selection can be calibrated by hyperparameter choices so that the ordering of models by posterior probability corresponds exactly to the ordering of models by each criterion of the form (3). The modal model of such a posterior will then correspond to the model that maximises (3). In § 3, the same hierarchical formulation is used to motivate a marginal maximum likelihood criterion that uses hyperparameter estimates based on the data rather than fixed choices. In § 4, we propose an alternative conditional maximum likelihood criterion based on maximizing rather than marginalising over the model space. The rapidly computable, closed form of this criterion is seen to provide an adaptive penalty that contains both a BIC type component that is $O(\log n)$ and an RIC type component that is similar to the modified RIC in (4). In § 5, we describe shrinkage estimators for the coefficients selected by our empirical Bayes criteria. In § 6, we report on simulation evaluations of our empirical Bayes criteria and compare them with AIC, C_p, BIC, RIC, modified RIC and a Cauchy prior modification of BIC proposed by Kass & Wasserman (1995). We consider various orthogonal and nonorthogonal set-ups, including the wavelet regression set-up where $X = I$ and $p = n$. In § 7, we conclude with a discussion of some recent related work.

2 Bayesian Calibration to Selection Criteria

We begin by showing that a particular class of priors on β_γ and γ may be calibrated so that the ordering of models by posterior probabilities $p(\gamma \mid Y)$ will be identical to the ordering of models by $SS_\gamma/\hat{\sigma}^2 - F\, q_\gamma$ in (3); to obtain the desired calibration, we treat σ as known and later set it equal to $\hat{\sigma}$. The priors we consider are of the form $p(\beta_\gamma, \gamma \mid c, w) = p(\beta_\gamma \mid \gamma, c)p(\gamma \mid w)$, where

$$p(\beta_\gamma \mid \gamma, c) = N_{q_\gamma}\{0, c\,\sigma^2(X_\gamma' X_\gamma)^{-1}\} \tag{5}$$

for $c > 0$, and

$$p(\gamma \mid w) = w^{q_\gamma}(1 - w)^{p - q_\gamma} \tag{6}$$

for $w \in (0, 1)$. Through its influence on the prior covariance matrix of β_γ in (5), the hyperparameter c controls the expected size of the nonzero coefficients of $\beta = (\beta_1, \ldots, \beta_p)'$. Under (6), the components of β are independently and identically nonzero with probability w, and so the hyperparameter w controls the expected proportion of such nonzero components.

An appealing aspect of (5) is its analytical tractability; it is of conjugate form, allowing analytical integration over β_γ for Bayesian marginalisation. Furthermore, conditional priors of the form (5) provide a coherent description of uncertainty in the sense that they are the conditional distributions of the nonzero components of β given γ when $\beta \sim N_p\{0, c\,\sigma^2(X'X)^{-1}\}$. The joint prior $p(\beta_\gamma, \gamma \mid c, w)$ corresponds to a reweighting of these conditional distributions according to $p(\gamma \mid w)$ in (6). As a result of their analytical and computational tractability, priors of the form (5) and (6) have often been used in Bayesian variable selection problems, typically with c and w set to fixed values. For example, this prior formulation was used by Smith & Kohn (1996) in a variable selection problem involving splines where they set $c \simeq 100$ and $w = 1/2$.

To reveal the connection between criteria of the form (3) and the posterior of γ under (5) and (6), we re-express the posterior as

$$\begin{aligned} p(\gamma \mid Y, c, w) &\propto p(\gamma \mid w)\, p(Y \mid \gamma, c) \\ &\propto w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{-\frac{Y'Y - SS_\gamma}{2\sigma^2} - \frac{SS_\gamma}{2\sigma^2(1+c)}\right\} \\ &\propto \exp\left[\frac{c}{2(1+c)}\{SS_\gamma/\sigma^2 - F(c,w)\,q_\gamma\}\right], \qquad (7) \end{aligned}$$

where

$$F(c,w) = \frac{1+c}{c}\left\{2\log\frac{1-w}{w} + \log(1+c)\right\}. \qquad (8)$$

These expressions reveal that, as a function of γ for fixed Y, $p(\gamma \mid Y, c, w)$ is increasing in

$$SS_\gamma/\sigma^2 - F(c,w)\,q_\gamma. \qquad (9)$$

The following is immediate.

THEOREM 1. *Under the prior obtained by* (5) *and* (6), *the posterior distribution is calibrated to* (9) *in the sense that for any two models* γ_1 *and* γ_2,

$$p(\gamma_1 \mid Y, c, w) > p(\gamma_2 \mid Y, c, w) \text{ if and only if}$$

$$SS_{\gamma_1}/\sigma^2 - F(c,w)\,q_{\gamma_1} > SS_{\gamma_2}/\sigma^2 - F(c,w)\,q_{\gamma_2},$$

where $F(c,w)$ *is defined by* (8). *Furthermore, the posterior mode occurs at the model* γ *for which* (9) *is maximised.*

If we compare (3) and (9), Theorem 1 shows that model selection via (3) with dimensionality penalty $F = F(c,w)$ is equivalent to model selection via posterior probabilities under the priors (5) and (6) with $\sigma^2 = \hat{\sigma}^2$. This correspondence between the two seemingly different approaches to variable selection provides additional insight and interpretability for users of either approach. In particular, when c and w are such that $F(c,w) = 2, \log n$ or $2\log p$, the highest posterior model will then correspond exactly to the best AIC/C_p, BIC or RIC models, respectively. Such equivalences may be seen as a refinement of the results of Smith & Spiegelhalter (1980), who showed an approximate correspondence between Bayes factors for nested normal-linear models under conjugate priors, and related selection criteria including AIC and BIC. Since posterior probabilities are monotone in (3), Theorem 1 provides additional justification for the common practice of maximising (3) within subsets of models in large problems where global maximisation is not computationally feasible. For such problems, Theorem 1 also reveals that it is possible to search for large values of (3) using Bayesian computational approaches that stochastically search for high posterior probability models; see George & McCulloch (1997).

Since c and w control the expected size and proportion of the nonzero components of β, the dependence of $F(c, w)$ on c and w in Theorem 1 provides an implicit connection between F and the profile of models for which its value may be appropriate. For example, $F(c, w) = 2, \log n$ and $2 \log p$ are obtained when $c \simeq 3.92, n$ and p^2 and $w = 1/2$. Such values are also obtained by other choices for (c, w) since $F(c, w)$ is increasing in c but decreasing in w. Indeed, $F(c, w)$ increases with the expected size of the nonzero coefficients, but decreases as the expected proportion of nonzero coefficients increases. We note in passing that $F(c, w)$ will be negative whenever $w > \{1 + (c + 1)^{\frac{1}{2}}\}^{-1} (> 1/2)$. In such cases, it follows from Theorem 1 that the highest posterior model will always be the full model. This corresponds to the prior information that the nonzero coefficients are many but small. As a result of the difficulty of distinguishing coefficients from noise in such a setting, the Bayesian formulation suggests using the full model and avoiding selection.

3 Empirical Bayes Estimation of c and w

Theorem 1 shows how choices for F in (3) correspond to choices for c and w, which in turn correspond to the expected size and proportion of the nonzero components of β. If subjective prior estimates of such model characteristics were available, these could be then used to guide the choice of c and w. For example, large c and small w would concentrate the prior on parsimonious models with large coefficients, and small c and large w would concentrate on saturated models with small coefficients.

In the absence of prior information, the arbitrary selection of c and w may tend to concentrate the prior away from the true underlying model, especially when p is large. For example, even the popular 'noninformative' choice $w = 1/2$ yielding $p(\gamma) \equiv 2^{-p}$ will concentrate on models with close to $p/2$ nonzero coefficients, which could be unsatisfactory when the true model was parsimonious or saturated. A default choice for c is also complicated because too small a value will yield a prior dominating the likelihood, and too large a value will tend to favour the null model. Furthermore, the choice of c should account for the sample size n, since increasing the data with a fixed choice of c would decrease the prior variance on the coefficients in (5). Indeed, by treating c explicitly as a function of n, Zellner (1986) and others have recommended choosing $c = O(n)$ in priors such as (5). In problems where $\mathrm{tr}(X_\gamma' X_\gamma) = O(n)$, this will prevent the prior from asymptotically dominating the likelihood.

When meaningful prior information about c and w is unavailable, as is usually the case in such a complicated context, perhaps the most reasonable strategy would be a fully Bayes approach that puts weak hyperprior distributions on c and w; for the general case where σ is unknown, it would also be sensible to put a weak prior on σ. Posterior computation of such a fully Bayes solution could, at least in principle, be obtained by Markov chain Monte Carlo methods similar to those described in George & McCulloch (1997). More precisely, one could simulate a Markov chain $(\gamma^{(1)}, c^{(1)}, w^{(1)}), (\gamma^{(2)}, c^{(2)}, w^{(2)}), \ldots$ that converges in distribution to $p(\gamma, c, w|Y)$. This could be accomplished using the Gibbs sampler or Metropolis-Hastings algorithms to make successive transitions from $\gamma^{(j)}$ to $\gamma^{(j+1)}$, and successive substitution sampling from $p(c \mid Y, \gamma, w)$ and $p(w \mid Y, \gamma, c)$. The potential drawback of this procedure is the computational limitation of visiting only a very small portion of the posterior when p is large. Although one might be able to use such methods to perform successfully a stochastic search for promising models (George & McCulloch, 1993, 1995, 1997) in such problems, it may not be feasible to get reliable estimates of c and w.

To avoid some of the difficulties of a fully Bayes approach, we propose an empirical Bayes approximation that uses the data to estimate c and w. Although such an approximation ignores the uncertainty of the estimates by treating them as known, as opposed to a fully Bayes marginalisation over c and w, it at least avoids using arbitrary choices of c and w which may be at odds with the data. Estimators of c and w for this purpose can be obtained by maximising the marginal likelihood of c and w under the prior (5) and (6), namely

$$\begin{aligned} L(c, w \mid Y) &\propto \sum_\gamma p(\gamma \mid w)\, p(Y \mid \gamma, c) \\ &\propto \sum_\gamma w^{q_\gamma} (1-w)^{p-q_\gamma} (1+c)^{-q_\gamma/2} \exp\left\{ \frac{c\, SS_\gamma}{2\sigma^2(1+c)} \right\}. \end{aligned} \tag{10}$$

Such maximum marginal likelihood estimators would correspond to posterior mode estimators under a fully Bayes formulation with independent uniform priors on c and w, a natural default choice.

Maximisation of (10) is, in general, computationally expensive when p is large because of the very large number of terms that must be summed. However, (10) simplifies considerably when X is orthogonal. In this case, if we let $t_i = b_i v_i / \sigma$ where v_i^2 is the ith diagonal element of $X'X$ and b_i is the ith component of $\hat{\beta} = (X'X)^{-1} X'Y$,

(10) reduces to

$$L(c, w \mid Y) \propto \prod_{i=1}^{p} \left\{(1-w)e^{-t_i^2/2} + w(1+c)^{-1/2}e^{-t_i^2/2(1+c)}\right\}. \tag{11}$$

Since only p distinct terms are involved, it can be substantially less expensive to compute (11) than (10). Even for moderately large p, it is feasible to use numerical methods to maximise (11).

The estimators $\hat{c}$ and $\hat{w}$ which maximise the marginal likelihood L in (10) or (11) can be used as prior inputs for an empirical Bayesian analysis with the priors (5) and (6). In particular, Theorem 1 shows that the model maximising the posterior $p(\gamma \mid Y, \hat{c}, \hat{w})$ based on these inputs is the γ that maximises

$$C_{\text{MML}} = SS_\gamma/\sigma^2 - F(\hat{c}, \hat{w})\, q_\gamma, \tag{12}$$

where

$$F(\hat{c}, \hat{w}) = \frac{1+\hat{c}}{\hat{c}} \left\{2\log\frac{1-\hat{w}}{\hat{w}} + \log(1+\hat{c})\right\}. \tag{13}$$

Thus, empirical Bayes selection is obtained by selecting the γ that maximises C_{MML}, which we refer to as the marginal maximum likelihood criterion. When σ^2 is unknown, as is usually the case in practise, we simply replace σ^2 in C_{MML} by an estimator $\hat{\sigma}^2$ such as one of those mentioned in § 1.

In contrast to criteria of the form (3), which penalise $SS_\gamma/\hat{\sigma}^2$ by Fq_γ, with F constant, C_{MML} uses an adaptive penalty $F(\hat{c}, \hat{w})$ that is implicitly based on the estimated distribution of the regression coefficients. Thus, C_{MML} may be preferable to using (3) with an arbitrary choice of F when little is known about c and w. This is supported by the simulation evaluations in § 6, which suggest that, compared to fixed choices of (3), selection using C_{MML} delivers excellent performance over a much wider portion of the model space.

4 A Conditional Empirical Bayes Approach

Although empirical Bayes selection using C_{MML} may be computationally more attractive than a fully Bayes approach, it can still be computationally overwhelming, especially when X is not orthogonal. Furthermore, neither C_{MML} nor fully Bayes solutions can be expressed in closed form, making it difficult to understand their relationship to alternative criteria. We now consider a further approximation to the fully Bayes solution; one that can be computed quickly and lends itself to revealing interpretation.

The idea behind our empirical Bayes derivation of C_{MML} was to approximate a fully Bayes solution by maximising $L(c, w \mid Y)$ rather than marginalising over c and

w. We now take this approximation one step further by eliminating the intermediate marginalisation over γ, and instead maximise the conditional 'likelihood'

$$\begin{aligned} L^*(c, w, \gamma \mid Y) &\propto p(\gamma \mid w) p(Y \mid \gamma, c) \\ &\propto w^{q_\gamma}(1-w)^{p-q_\gamma}(1+c)^{-q_\gamma/2} \exp\left\{\frac{c\, SS_\gamma}{2\sigma^2(1+c)}\right\}, \end{aligned} \tag{14}$$

which is also proportional to the posterior (7). Strictly speaking, L^* is not a likelihood because γ is an unknown intermediate quantity that is more like a missing value than a parameter. Although marginalisation is preferable to maximisation, we shall see that considerable insight and computational simplicity is obtained by this conditional empirical Bayes approach.

Before proceeding, we remark that maximising $L^*(c, w, \gamma \mid Y)$ is equivalent to maximising the largest component of $L(c, w|Y)$, so that, when L is strongly dominated by a single component, the conditional and marginal empirical Bayes approaches should yield similar answers. The nature of this approximation is further revealed by noting that, when X is orthogonal, L^* can be reexpressed as

$$L^*(c, w, \gamma \mid Y) = \left\{\prod_{x_i \notin \gamma} (1-w) e^{-t_i^2/2}\right\} \left\{\prod_{x_i \in \gamma} w(1+c)^{-1/2} e^{-t_i^2/2(1+c)}\right\}. \tag{15}$$

In this case, maximising L^* will be equivalent to maximising the dominant part of each term of L in (11).

Conditionally on γ, the estimators of c and w that maximise L^* are easily seen to be

$$\hat{c}_\gamma = \left(SS_\gamma/\sigma^2 q_\gamma - 1\right)_+ , \tag{16}$$

where $(\cdot)_+$ is the positive-part function, and

$$\hat{w}_\gamma = q_\gamma / p\,. \tag{17}$$

Inserting these into the posterior (7), or equivalently into L^*, and taking the logarithm, shows that the posterior $p(\gamma \mid Y, \hat{c}_\gamma, \hat{w}_\gamma)$ is maximised by the γ that maximises

$$C_{\text{CML}} = SS_\gamma/\sigma^2 - B(SS_\gamma/\sigma^2) - R(q_\gamma) \tag{18}$$

where

$$B(SS_\gamma/\sigma^2) = q_\gamma \left\{1 + \log_+(SS_\gamma/\sigma^2 q_\gamma)\right\}, \tag{19}$$

$\log_+(\cdot)$ is the positive part of $\log(\cdot)$, and

$$R(q_\gamma) = -2\left\{(p - q_\gamma)\log(p - q_\gamma) + q_\gamma \log q_\gamma\right\}. \tag{20}$$

Thus, selecting the γ that maximises C_{CML}, which we refer to as the conditional maximum likelihood criterion, provides an approximate empirical Bayes alternative to selection based on C_{MML}. As with C_{MML}, when σ^2 is unknown, we simply replace σ^2 in C_{CML} by an estimator $\hat{\sigma}^2$ such as one of those mentioned in § 1.

In contrast to the adaptive penalty of C_{MML}, C_{CML} has an adaptive penalty that can be expressed in closed form and is rapidly computable, even when X is not orthogonal. We have purposely expressed this penalty as the sum of two components $B(SS_\gamma/\sigma^2) + R(q_\gamma)$ to highlight its interesting interpretability. As we now show, this penalty acts like a combination of a modified BIC penalty $F = \log n$ and a modified RIC penalty $F = 2\log p$. Insofar as maximising C_{CML} approximates maximising C_{MML}, as is supported by the simulations in § 6, these interpretations should at least roughly explain the behaviour of the C_{MML} penalty $F(\hat{c}, \hat{w})$ in (13).

The $B(SS_\gamma/\sigma^2)$ component of C_{CML} penalises the addition of a variable to the model by $1 + \log_+(SS_\gamma/\sigma^2 q_\gamma)$. Assuming the diagonal elements of $X'X$ are $O(n)$, $1 + \log_+(SS_\gamma/\sigma^2 q_\gamma)$ will be asymptotically equivalent to the BIC penalty $\log n$ whenever β_γ has at least one nonzero component; in such cases, $SS_\gamma = O(n)$ and hence $1 + \log_+(SS_\gamma/\sigma^2 q_\gamma) = O(\log n)$. Furthermore, the values of C_{CML} for such nonnull β_γ will asymptotically dominate the other values of C_{CML}. Thus, whenever β has at least one nonzero component, the dominant terms of C_{CML} will asymptotically differ from BIC only by terms of constant order, and so the model maximising C_{CML} will be consistent for γ as $n \to \infty$; see Kass & Wasserman (1995). These statements also remain true when σ^2 has been replaced by a consistent estimator $\hat{\sigma}^2$. It is interesting to note that for such nonnull β, the implicit conditional maximum likelihood estimator of c in (16) is $O(n)$, thereby preventing the implicitly estimated prior from asymptotically dominating the likelihood as discussed early in § 3.

When $\beta \equiv 0$, the distribution of values of C_{CML} will not depend on n, and although the model maximising C_{CML} will be small with substantial probability, it will also always have a positive probability of being incorrect. In spite of the consistency of the least squares coefficient estimators, this behaviour reflects the fact that for any n there will always be a neighbourhood of nonzero β components that are difficult to distinguish from zero. Thus, in the spirit of the recommendation of Mallows (1973) for C_p, one could treat C_{CML} as a diagnostic tool which either picks up a strong signal or reflects the fact that the data are inconclusive. In the latter case, one could simply decide to choose the null model on the basis of practical considerations. An automatic implementation that tends to accomplish this is proposed below.

The $R(q_\gamma)$ component of C_{CML} is proportional to the entropy of a Bernoulli distribution with $\pi = q_\gamma/p$. It penalises the addition of a variable to the model by approximately

$$2\log\frac{p - q_\gamma + 1}{q_\gamma}; \tag{21}$$

this follows from the approximation $R(q) - R(q-1) \simeq 2\log\{(p-q+1)/q\}$, which is obtained using $\log x + 1 \simeq x \log x - (x-1)\log(x-1)$. The quantity in (21) is identical to the RIC penalty when $q_\gamma = 1$, which corresponds to the addition of a variable to the null model. However, in contrast to RIC, the incremental penalty (21) decreases as more variables are added. The adjustment is similar to the incremental penalty $2\log(p/q_\gamma)$ in (4), which is approximately the expected size of the qth largest squared t-statistic under the null model. However, (21) is less than $2\log(p/q_\gamma)$, especially for q_γ large. In fact, after $p/2$ variables have been included, (21) becomes negative. When this happens and SS_γ/q_γ is small enough, the total penalty $B(SS_\gamma/\sigma^2) + R(q_\gamma)$ will actually reward, rather than penalise, the inclusion of more variables. Although surprising at first, this behaviour corresponds to the behaviour of the posterior (7) when $F(c, w)$ is negative, as discussed at the end of § 2. Indeed, when the data suggest that the nonzero coefficients are many but small, C_{CML} will favour including them. Such behaviour will also occur with C_{MML} when its penalty $F(\hat{c}, \hat{w})$ in (13) is negative. As is borne out by the simulations in § 6, this behaviour allows both C_{MML} and C_{CML} to perform very well when the true q_γ is large.

Since it so easily computed, C_{CML} is an attractive alternative to C_{MML} for practical implementation. For this purpose, however, we have found that one can do better than simply choosing the maximum C_{CML} model. Our preliminary simulation investigations revealed that, unless the true coefficients are large, C_{CML} often tends to be bimodal over the model space, with one mode closer to the true model and the other at the completely saturated model, unless the true model happens to be the saturated model. A direct consequence of the bimodal $R(q_\gamma)$ penalty, this behaviour stems from the fact that the likelihood does not distinguish well between models with small c and small w and models with even smaller c but large w. Since the largest mode will sometimes occur at the incorrect saturated model, the remedy we propose is simply to choose the more parsimonious of the modal models. This modification requires little additional computational effort and appears to improve performance substantially with respect to predictive loss (26) when β is equal or close to zero.

As with criteria of the form (3), full implementation of selection using C_{CML} requires exhaustive computation of all values of SS_γ; branch and bound strategies (Furnival & Wilson 1974) may help to lessen this burden. However, when p is so large that exhaustive computation is not feasible, C_{CML} can still be useful because it will at least identify local posterior modes within restricted subsets of models. Indeed, C_{CML} can still be effective with heuristic strategies such as stepwise search, as is borne out by the simulations in § 6.3. It may also be fruitful to use Bayesian stochastic search methods to identify promising subsets of models for consideration; see George & McCulloch (1997).

5 Estimating β_γ After Selection

After a model has been selected by either C_{MML} or C_{CML}, it will generally be of interest to estimate the values of the selected coefficients. A common, but often inferior, strategy is to ignore the fact that selection was used and simply estimate the selected coefficients by least squares. Such a strategy ignores selection bias and leads to overestimates. Alternatives that moderate this bias arise naturally from our Bayesian formulation.

Under the priors (5) and (6), the conditional Bayes estimator of β_γ given γ is

$$E(\beta_\gamma \mid Y, c, \gamma) = \frac{c}{1+c}\,\hat{\beta}_\gamma, \tag{22}$$

where $\hat{\beta}_\gamma \equiv (X'_\gamma X_\gamma)^{-1} X'_\gamma Y$. Thus, after γ has been selected by either C_{MML} or C_{CML}, the corresponding estimator of β_γ is obtained by substituting the appropriate estimate of c. The conditional estimator based on C_{MML} substitutes the numerically computed maximum marginal likelihood estimator $\hat{c}$ to yield

$$\hat{\beta}_\gamma^{\text{MML}} = \frac{\hat{c}}{1+\hat{c}}\,\hat{\beta}_\gamma. \tag{23}$$

The conditional estimator based on C_{CML} substitutes the estimator $\hat{c}_\gamma$ in (16) to yield the Stein-like estimator (Stein, 1981)

$$\hat{\beta}_\gamma^{\text{CML}} = \left(1 - \frac{\sigma^2 q_\gamma}{SS_\gamma}\right)_+ \hat{\beta}_\gamma. \tag{24}$$

When σ^2 is unknown, we simply replace σ^2 in (24) by the same estimator $\hat{\sigma}^2$ used in C_{CML}. It is trivial to compute $\hat{\beta}_\gamma^{\text{MML}}$ or $\hat{\beta}_\gamma^{\text{CML}}$ once γ has been selected. Although neither of these estimators will be minimax under predictive loss, the simulation results in § 6 reveal that they obtain consistent improvements over the use of selection followed by least squares estimation.

If the goal of analysis is exclusively prediction, one may prefer Bayes estimators which are not conditional on any γ and correspond to model averaging. Under the priors (5) and (6), the unconditional Bayes estimator of $\beta = (\beta_1, \ldots, \beta_p)'$ is

$$E(\beta \mid Y, c, w) = \sum_\gamma p(\gamma \mid Y, w) E(\beta \mid Y, c, \gamma). \tag{25}$$

The corresponding marginal or conditional maximum likelihood estimators would then be obtained by substituting the corresponding estimators for c and w into (25). Although such estimators are likely to provide better predictions than the conditional estimators, they would also yield nonzero estimates of all p coefficients, regardless of how many of them are actually zero. The attraction of the conditional estimators $\hat{\beta}_\gamma^{\text{MML}}$

and $\hat{\beta}_\gamma^{\text{CML}}$ is that they will usually provide a parsimonious representation of simple models, thereby facilitating interpretation and understanding. They also are much less expensive to compute, especially when p is large. However, when X is orthogonal, estimators of the form (25) can be simplified considerably, from an average over 2^p models to p separate, rapidly computable, closed-form estimates of the individual components; see Clyde, Parmigiani & Vidakovic (1998).

6 Simulation Evaluations

6.1 Preamble

In this section we report on simulation investigations of the performance potential of selection using C_{MML} in (12) and C_{CML} in (18), and the conditional shrinkage estimators $\hat{\beta}_\gamma^{\text{MML}}$ in (23) and $\hat{\beta}_\gamma^{\text{CML}}$ in (24). For comparison, we included the three fixed penalty selection criteria based on (3): AIC/C$_p$, BIC and RIC. We also included two additional criteria with varying penalties: modified RIC in (4), and the criterion $SS_\gamma/\sigma^2 - q_\gamma \log n - \log r(q_\gamma)$, where $r(q_\gamma) = \pi^{\frac{1}{2}}/\{2^{(p-q_\gamma)/2}\Gamma(\frac{p-q_\gamma+1}{2})\}$, which we refer to as Cauchy BIC. Showing that BIC is a close asymptotic Bayes factor approximation under a Normal unit information prior, Kass and Wasserman (1995) proposed what we call Cauchy BIC as a Bayes factor approximation under a Cauchy unit information prior, thereby generalising the Cauchy default proposal of Jeffreys (1961). Pauler (1998) discusses the close relationship between Cauchy BIC and other natural default Bayes factor approximations such as those proposed by Zellner & Siow (1980), O'Hagan (1995) and Berger & Pericchi (1996). For reasons of space and focus, we did not include these criteria here, but plan to report elsewhere on comparisons with these.

For various choices of X and β described below, we simulated data from the regression model (1) with $\sigma^2 = 1$; rather than introduce more noise through an independent estimate of σ^2, we simply assumed that $\sigma^2 = 1$ was known. At each iteration, and for each selection criterion, we summarised the disparity between the selected model $\hat{\gamma}$ and the correct γ underlying β by the predictive loss

$$L\{\beta, \hat{\beta}(\hat{\gamma})\} \equiv \{X\hat{\beta}(\hat{\gamma}) - X\beta\}'\{X\hat{\beta}(\hat{\gamma}) - X\beta\} \tag{26}$$

with $\hat{\beta}_i(\hat{\gamma})$ equal to the least squares estimate of β_i when $\hat{\gamma}$ includes x_i and 0 otherwise. Although other more complicated evaluations could have been used, such as frequencies of incorrect inclusions and exclusions, we chose (26) primarily because it provides a natural scalar summary of the disparity between $\hat{\gamma}$ and γ. We also used (26) to evaluate $\hat{\beta}_\gamma^{\text{MML}}$ and $\hat{\beta}_\gamma^{\text{CML}}$, with $\hat{\beta}_i(\hat{\gamma})$ equal to the corresponding coordinate estimate when $\hat{\gamma}$ includes x_i and 0 otherwise. We then summarised overall performance, in Figures 1 and 2 below, by the average loss over repeated iterations in each of our various set-ups. The numerical simulation summaries are available on the website http://bevo2.bus.utexas.edu/GeorgeE/.

6.2 A Simple Orthogonal Set-up

Here, we let $X = I$, reducing (1) to $Y = \beta + \epsilon$, so that the variable selection problem becomes one of identifying the nonzero components of a multivariate normal mean. Note that here $p = n$. This problem occurs naturally in the context of wavelet regression; see Donoho & Johnstone (1994).

To simulate a value of Y here, we fixed values of c and $q \leq p$, generated the first q components of β as $\beta_1, \ldots, \beta_q \sim N(0, c)$ independently, set $\beta_{q+1}, \ldots, \beta_p \equiv 0$ and then added $\epsilon \sim N_p(0, I)$ to β. Each selection criterion was then applied to Y, and its loss (26) was evaluated using $\hat{\beta}_i(\hat{\gamma}) = y_i I[x_i \in \hat{\gamma}]$, the least squares estimate of β_i under the selected model $\hat{\gamma}$. The loss for the shrinkage estimators was evaluated by using the shrinkage estimates instead of the least squares estimates. With $p = n = 1000$, this simulation was repeated 2500 times for $c = 5$ and 25 and $q = 0$, 10, 25, 50, 100, 200, 300, 400, 500, 750 and 1000. For each pair of c and q, the loss for each procedure was averaged over the corresponding 2500 repetitions. To help gauge these average losses, note that here the expected loss of the full least squares estimator of β is always $p = 1000$, and the expected loss of the 'oracle' criterion, which selects the correct γ and then uses least squares, is q.

It should be mentioned that the orthogonality of X was crucial for the computational feasibility of these simulations. It allowed us to avoid evaluation of all 2^{1000} models, and instead restrict attention to a sequence of 1000 easily computed SS_γ values. It also allowed us to calculate C_{MML} by numerical maximisation of the simplified likelihood (11). This maximisation was performed with the constrained maximisation routine CONSTR in Matlab; to improve numerical stability over so many repetitions, we imposed the additional constraint $c \geq 0.5$.

Figures 1(a) and 1(b) present the results when $c = 25$. If we focus first on the fixed penalty criteria AIC/C$_p$, BIC and RIC, which here use penalties of $F = 2$, log 1000 $\simeq 6.9$, and 2 log 1000 $\simeq 13.8$, the performance of each of these criteria deteriorates linearly as q increases. The larger penalty values yield much better performance when q is small, but then deteriorate much more rapidly as q increases, resulting in a relative performance reversal when q is large. Turning to the other criteria, we see that the performance of Cauchy BIC is nearly identical to that of RIC, being very good for small q but very poor for large q. Apparently variation of its penalty function does not have much of an effect. In contrast to this, variation of the modified RIC penalty has a much more pronounced effect; like RIC, its performance is also very good when q is small, but deteriorates much more slowly and nonlinearly as q increases. However, the most striking performance is that of C_{MML} which appears to emulate adaptively the performance of the best fixed penalty criterion for each q. Indeed,

C_{MML} appears to dominate all the other criteria at every value of q. Apparently, its adaptive dimensionality penalty adapts very effectively to the model information contained in the data. Although C_{CML} did not perform quite as well for all q, it was excellent for small values of q, and offered similar adaptive advantages over the fixed penalty rules. Note that modified RIC, C_{MML} and C_{CML} all delivered nearly perfect performance when $q = 1000$, in sharp contrast to the other criteria.

Figure 1(c) presents the performance results when $c = 5$, which reduced the standard deviation of the coefficient distribution from 5 above to about 2.24. The only substantial changes in relative performance among the selection criteria occurred with modified RIC and C_{CML}. The performance of modified RIC became much closer to that of C_{MML}, which was still dominant. However, the performance of C_{CML} was not quite as good, and it seemed to deteriorate somewhat for larger values of q. This may in part be an effect of our strategy, discussed towards the end of § 4, of choosing the more parsimonious modal C_{CML} model in the presence of bimodality.

Finally, for every pair of c and q considered above, both $\hat{\beta}_\gamma^{\text{MML}}$ and $\hat{\beta}_\gamma^{\text{CML}}$ provided improved performance over their corresponding C_{MML} and C_{CML} least squares estimators. It was interesting that the $\hat{\beta}_\gamma^{\text{MML}}$ improvements were always greater than the $\hat{\beta}_\gamma^{\text{CML}}$ improvements, even though the C_{MML} performance was better to begin with. The median improvements by $\hat{\beta}_\gamma^{\text{MML}}$ and $\hat{\beta}_\gamma^{\text{CML}}$ were 3.4 % and 2.1 % when $c = 25$ and were 15.8 % and 3.4 % when $c = 5$. The largest improvements occurred when $q = 0$ and were 53.8 % by $\hat{\beta}_\gamma^{\text{MML}}$ and 10.0 % by $\hat{\beta}_\gamma^{\text{CML}}$.

6.3 Correlated Predictors and Fixed β Values

We next compared the performance of the various criteria in problems with correlated predictors for fixed values of β. Aspects of the following set-up were motivated by the simulation set-up in Breiman (1992). The n rows of X were independently simulated from a $N_p(0, \Sigma)$ distribution where the ijth element of Σ was $\rho^{|i-j|}$. We carried this out with $n = 200$, $p = 50$ and $\rho = -0.5, 0.0, 0.5$. We considered 11 choices for β, each consisting of five replications of $(\beta_1, \ldots, \beta_{10})'$, i.e. $\beta_i = \beta_{i-10}$ for $i = 11, \ldots, 50$. For $k = 0, 1, \ldots, 10$, our kth choice of $(\beta_1, \ldots, \beta_{10})$ consisted of k adjacent values of B_k centred around β_5 and zeroes otherwise. This led to 11 choices of β, each consisting of 0, 5, ..., 45 and 50 nonzero components respectively. Except for $k = 0$, the values of B_k were then chosen to yield a theoretical $R^2 = \beta' X' X \beta / (\beta' X' X \beta + 200)$ equal to 0.75. Note that B_k decreases as k increases. For each choice of ρ, X was held fixed while Y was simulated from $Y = X\beta + \epsilon$ where $\epsilon \sim N_{200}(0, I)$, for each of the 11 choices of β.

Since X was not orthogonal, because of both sampling variation and ρ, it was impractical to evaluate SS_γ for all 2^{50} models. This ruled out not only consideration of C_{MML} but also evaluation of the other criteria over the entire model space. Instead, we considered the common practice of applying the selection criteria to the subset of models obtained by a stepwise method. For each simulated Y, we simply used each criterion to select a model from one of the 51 models, 50 plus the null model, visited by forward selection. For each pair of ρ and β, this simulation was repeated 2500 times and the loss (26) for each procedure was averaged over the repetitions. To help gauge the average loss values, note that here the expected loss of the full least squares estimator of β is always $p = 50$, and the expected loss of the 'oracle' criterion, which selects the correct γ and then uses least squares, is $5k$, the number of nonzero components of our kth choice of β.

Figure 2(a) presents the results when $\rho = 0.0$. The first thing to note is that, overall, the results are qualitatively very similar to those in the previous simulations. The performances of AIC/C$_p$, BIC and RIC, which here use penalties of $F = 2$, log $200 \simeq 5.3$ and 2 log $50 \simeq 7.8$, again deteriorate as $5k$ increases. The larger penalty values yield much better performance when $5k$ is small, but then deteriorate much more rapidly, resulting in a relative performance reversal when $5k$ is large. The nonlinearity of the deterioration here is apparently caused by the decreasing size B_k of the coefficients. Cauchy BIC is again very similar to RIC in performance. Modified RIC again performs very well when $5k = 0$, deteriorating rapidly then slowly, levelling off to an average loss of 50 when $5k = 35$. C_{CML} is once again excellent for small and large values of $5k$, and only slightly worse than the performance of the best criteria at intermediate values.

Figure 2(b) presents the results when $\rho = 0.5$, inducing positive correlation between the predictors. With the exception of C_{CML}, the relative performances of the criteria remained the same, although the performance deterioration occured more rapidly with performance reversals occuring at smaller values of $5k$. Although C_{CML} was still excellent for small $5k$, its performance deteriorated more rapidly, levelling off at a larger average loss. However, it was only worse than AIC/C$_p$ and modified RIC at large $5k$, both of which it dominated substantially at small $5k$. Note that, when predictors with nonzero coefficients are left out of the model, the effect of the positive correlation is to inflate the coefficients of the remaining predictors. As a result, all the criteria will tend to bias selection towards parsimonious models because of their increased explanatory power. This may also lead stepwise selection to mistake noise for signal when including variables, further exacerbating this bias. Apparently this bias has a more severe effect on the performance of C_{CML}.

Figure 2(c) presents the results when ρ = -0.5, inducing negative correlation between the predictors. The relative performance of the fixed penalty criteria and Cauchy BIC again remained the same except that here performance deterioration occured more slowly, resulting in performance reversals at larger values of $5k$. The negative correlation here has the opposite effect to that of positive correlation, tending to bias selection towards saturated models. Surprisingly, the performance of modified RIC and C_{CML} was very similar to their performance in the $\rho = 0.0$ case. C_{CML} was again superior at the small and large values of $5k$, but because of the relatively slow deterioration, was modestly worse than the fixed penalty criteria and Cauchy BIC at intermediate values of $5k$.

Finally, for every pair of ρ and β considered above, $\hat{\beta}_{\gamma}^{\text{CML}}$ provided improved performance over the C_{CML} least squares estimator. The median improvements were 6.8 %, 3.7 % and 6.9 % for $\rho = 0.0, 0.5$ and -0.5, respectively. The largest improvements occurred when $\beta \equiv 0$ and were 15.4 %, 20.5% and 21.0 % for $\rho = 0.0, 0.5$ and -0.5, respectively.

7 Some Recent Related Work

During the course of revising this article, some related work appeared in the context of orthogonal nonparametric wavelet regression, namely a University of Bristol technical report by I. M. Johnstone and B. W. Silverman and Clyde & George (1999, 2000). These articles provide further results concerning the marginal and conditional maximum likelihood methods, and were motivated in part by a preliminary version of this article which appeared as George & Foster (1998).

Johnstone and Silverman proposed an effective EM algorithm for computing marginal maximum likelihood hyperparameter estimates, and showed that posterior median estimators based on these and on MAD estimates of σ^2, compare favourably with the BayesThresh method of Abramovich, Sapatinas & Silverman (1998). Noting that such hyperparameter estimates are consistent by classical maximum likelihood theory, they also showed that conditional maximum likelihood estimates of c and w are asymptotically biased, and that this bias can be severe when w is small.

The two articles by Clyde and George compared the performance of conditional and unconditional shrinkage estimators based on marginal and conditional maximum likelihood using MAD estimators of σ^2, and using marginal maximum likelihood estimates of σ^2 obtained by an EM algorithm. The performance of these empirical Bayes shrinkage estimators was seen to compare favourably with alternative fixed penalty methods. Although the bias of the conditional maximum likelihood hyperparameter estimators was also noted, the effect of this bias did not appear to denigrate seriously performance. This may be partially explained by the fact that the biases, c high and w low, have an offsetting effect in the posterior as revealed by the fact that $F(c, w)$ in (8) is increasing in c but decreasing in w. Examples of explicit graphical comparisons that reveal similarities between the marginal and conditional likelihoods were also presented by Clyde & George (1999). The marginal maximum likelihood approach was also extended to models with heavy tailed error distributions by Clyde & George (2000).

Acknowledgment

The authors are grateful to Susie Bayarri, Jim Berger, Hugh Chipman, Merlise Clyde, Petros Dellaportas, Guido Consonni, Sir David Cox, Phil Dawid, David Donoho, Alan Gelfand, Wally Gilks, Mark Hansen, Iain Johnstone, Colin Mallows, Rob McCulloch, Donna Pauler, Bernard Silverman, the Editor, the Associate Editor and the referees for many helpful suggestions. This work was supported by the National Science Foundation, the Texas Higher Education Coordinating Board Advanced Research Program, the University of Pennsylvania and the University of Texas.

References

Abramovich, F., Sapatinas, T., & Silverman, B.W. (1998). Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B **60**, 725-49.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. 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. R. Statist. Soc.* B **57**, 289-300.

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

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J. Am. Statist. Assoc.* **87**, 738-54.

Clyde, M.A., DeSimone, H. & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Am. Statist. Assoc.* **91**, 1197-1208.

Clyde, M., Parmigiani, G. & Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391-402.

Clyde, M. & George, E.I. (1999) Empirical Bayes estimation in wavelet nonparametric Regression. In *Bayesian Inference in Wavelet Based Models*, Ed. P. Muller and B. Vidakovic, pp. 309-22. New York: Springer-Verlag.

Clyde, M. & George, E.I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B **62**, (to appear).

Donoho, D.L., & Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-56.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with Discussion). *J. R. Statist. Soc.* B **57**, 301-69.

Foster, D.P., & George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947-75.

Foster, D.P., & Stine, R.A. (1999). Local asymptotic coding. *IEEE Transactions on Information Theory* **45**, 1289 - 1293.

Furnival, G.M. & Wilson, R.W. (1974) Regression by leaps and bounds. *Technometrics* **16**, 499-511.

Garthwaite, P. H. & Dickey, J.M. (1996). Quantifying and using expert opinion for variable-selection problems in regression (with discussion). *Chemomet. Intel. Lab. Syst.* **35**, 1-34.

George, E.I. & Foster, D.P. (1998). Empirical Bayes variable selection (with Discussion). In *Proceedings of the Workshop on Model Selection, Special Issue of Rassegna di Metodi Statistici ed Applicazioni*, Ed. W. Racugno, pp. 79-108. Bologna: Pitagora Editrice.

George, E.I., & McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881-89.

George, E.I., & McCulloch, R.E. (1995). Stochastic search variable selection. In *Practical Markov Chain Monte Carlo in Practice*, Ed. W.R. Gilks, S. Richardson and D.J. Spiegelhalter, pp. 203-14. London: Chapman and Hall.

George, E.I., & McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339-73.

Hansen, M. H. and Yu, B. (1999). Bridging AIC and BIC: An MDL Model Selection Criterion. In *Proceedings of the IT Workshop on Detection, Estimation, Classification and Imaging.* Santa Fe, NM.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Assoc.* **90**, 928-34.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-76.

Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression (with Discussion). *J. Am. Statist. Assoc.* **83**, 1023-36.

O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with Discussion). *J. R. Statist. Soc.* 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., Madigan, D. M., & Hoeting, J., (1997). Bayesian model averaging for linear regression models. *J. Am. Statist. Assoc.* **92**, 179-91.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Economet.* **75**, 317-44.

Smith, A.F.M. & Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models. *J. R. Statist. Soc.* B **42**, 213-20.

Stein, C. (1981). Estimation of a multivariate normal mean. *Ann. Statist.* **9**, 1135-51.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finietti*, Ed. P.K. Goel and A. Zellner, pp 233-43. Amsterdam:North-Holland.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics Proceedings of the First International Meeting Held in Valencia (Spain)*, Ed. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, pp. 585-603. Valencia: University of Valencia Press.

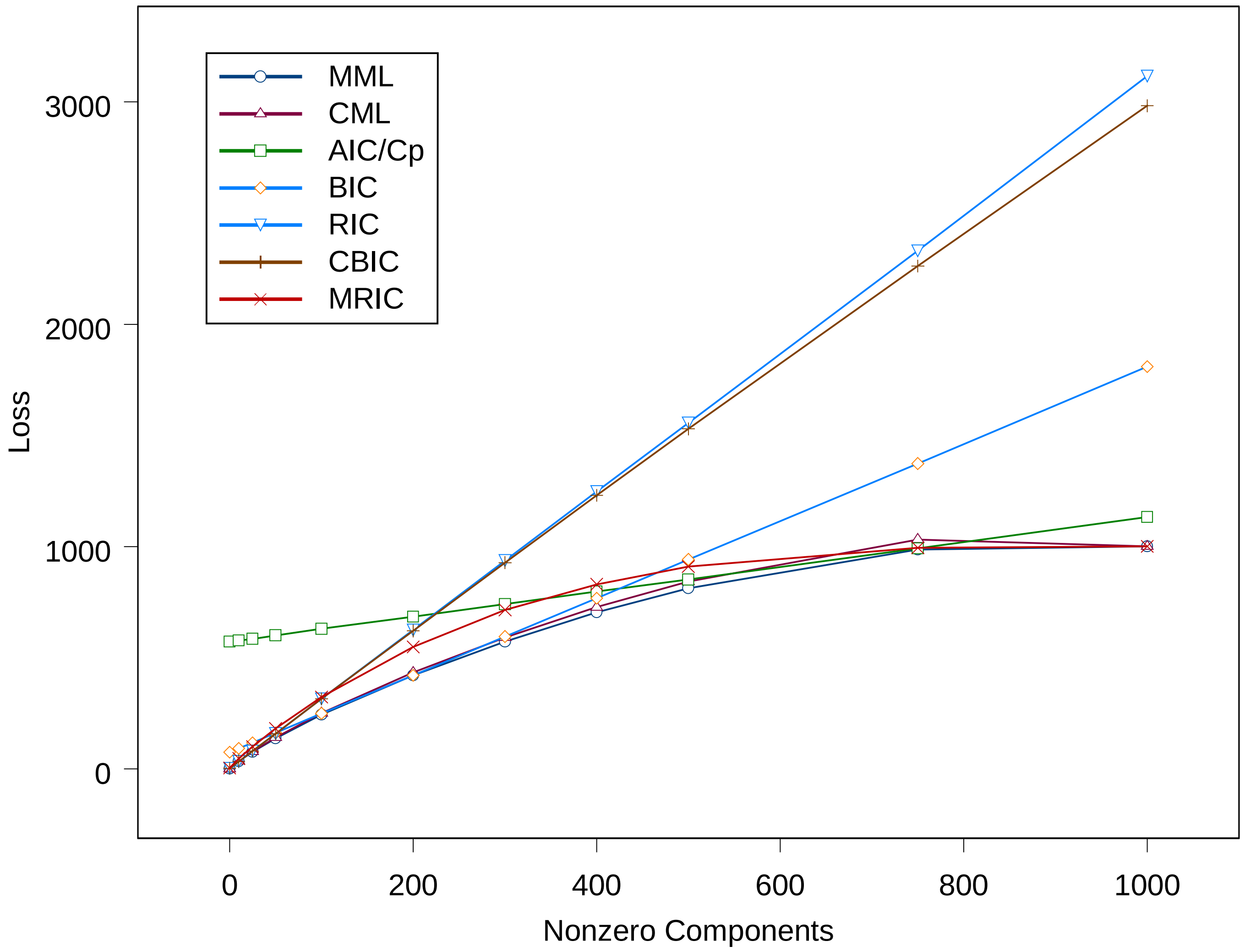

Figure 1(a). The average loss of the selection procedures when c = 25 and the number of nonzero components q = 0,10,25,50,100,200, 300, 400, 500, 750, 1000. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

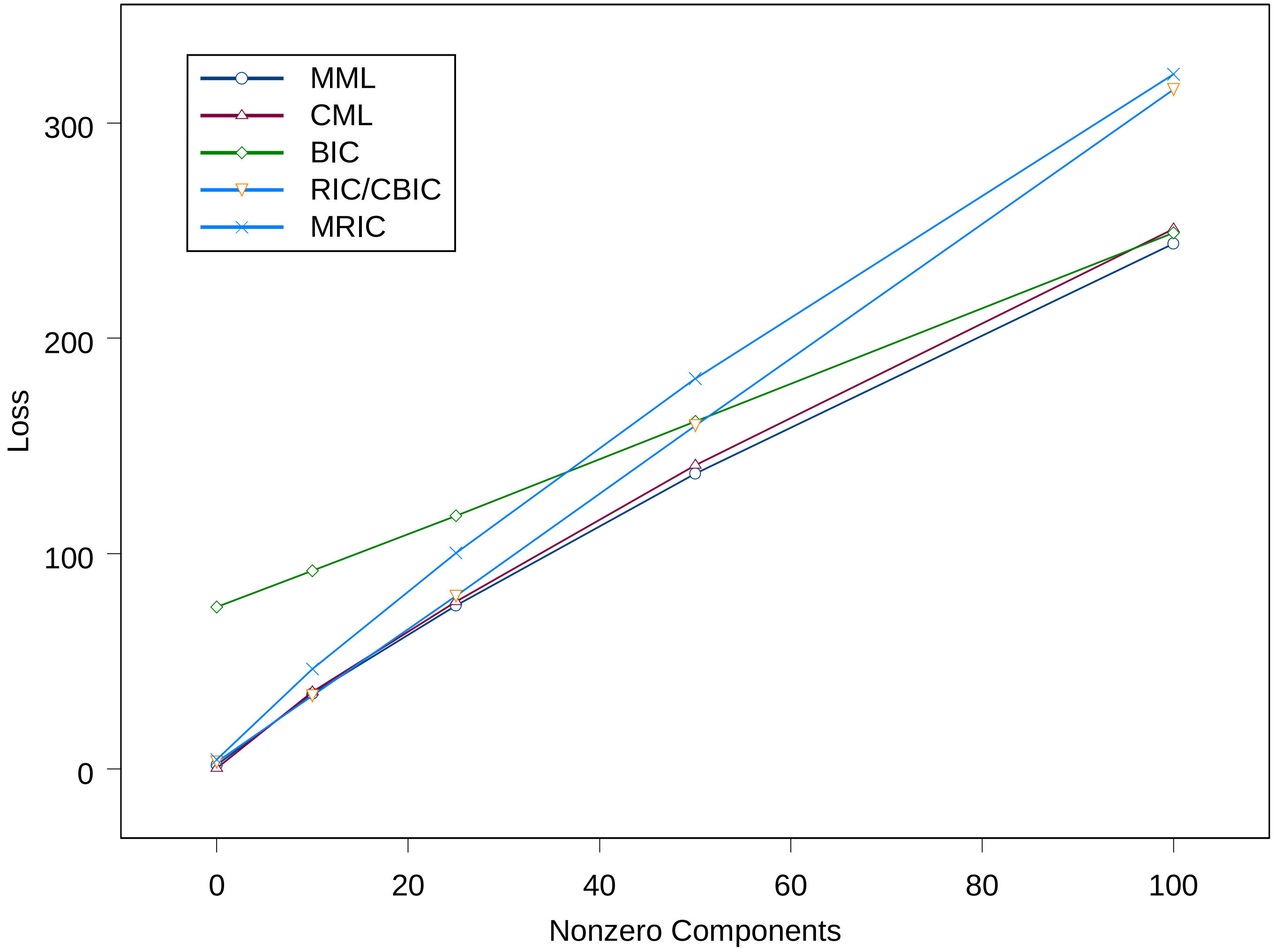

Figure 1(b). The average loss of the selection procedures when c = 25 and the number of nonzero components q = 0,10,25,50,100. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC. RIC and CBIC are virtually identical here and so have been plotted together.

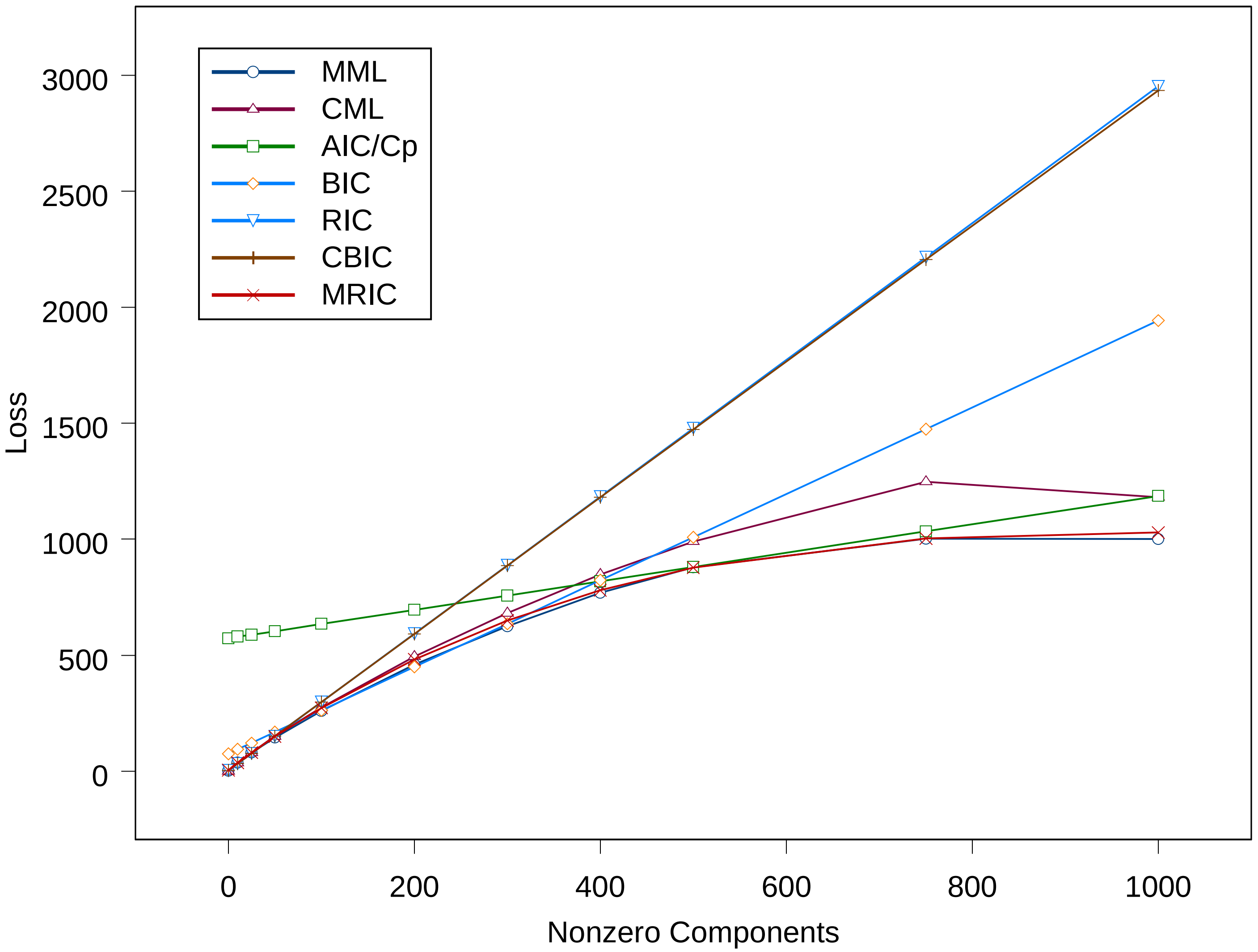

Figure 1(c). The average loss of the selection procedures when *c* = 5 and the number of nonzero components *q* = 0,10,25,50,100,200, 300, 400, 500, 750, 1000. We denote C_{MML} by MML, C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

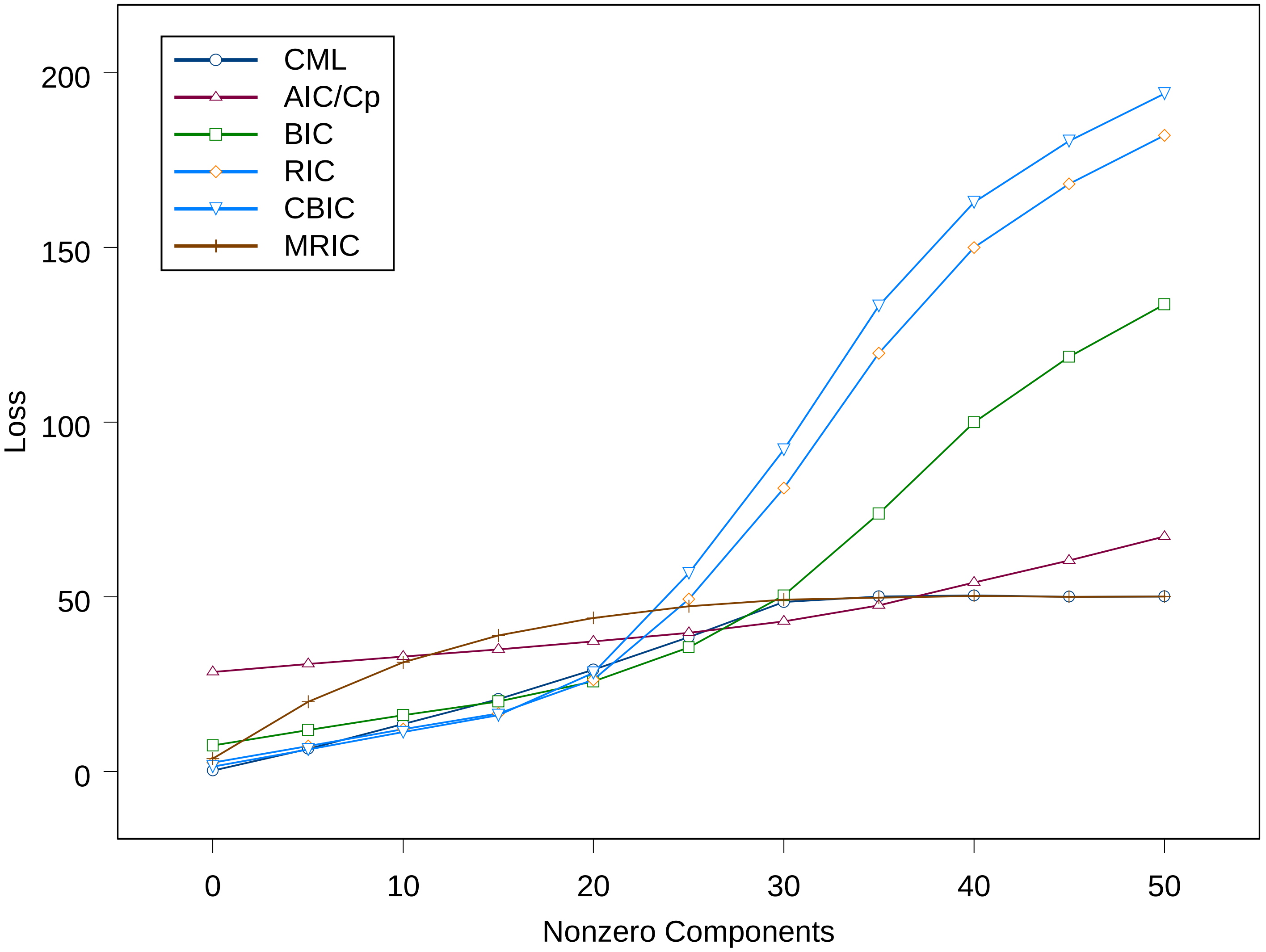

Figure 2(a). The average loss of the selection procedures when $\rho = 0.0$ for 11 choices of β with 0, 5, 10,...,50 nonzero components, respectively. We denote C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

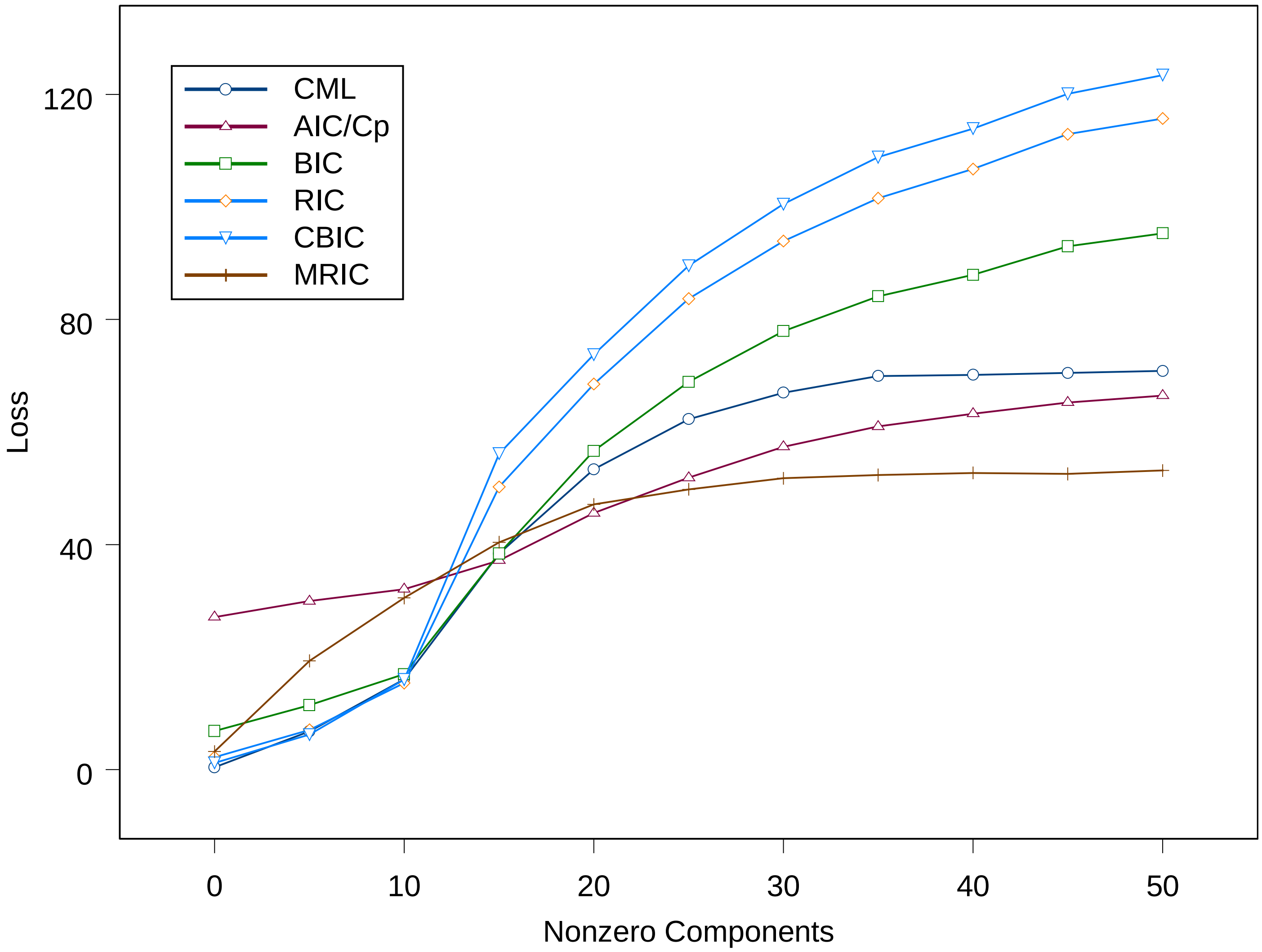

Figure 2(b). The average loss of the selection procedures when $\rho = 0.5$ for 11 choices of β with 0, 5, 10,...,50 nonzero components, respectively. We denote C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.

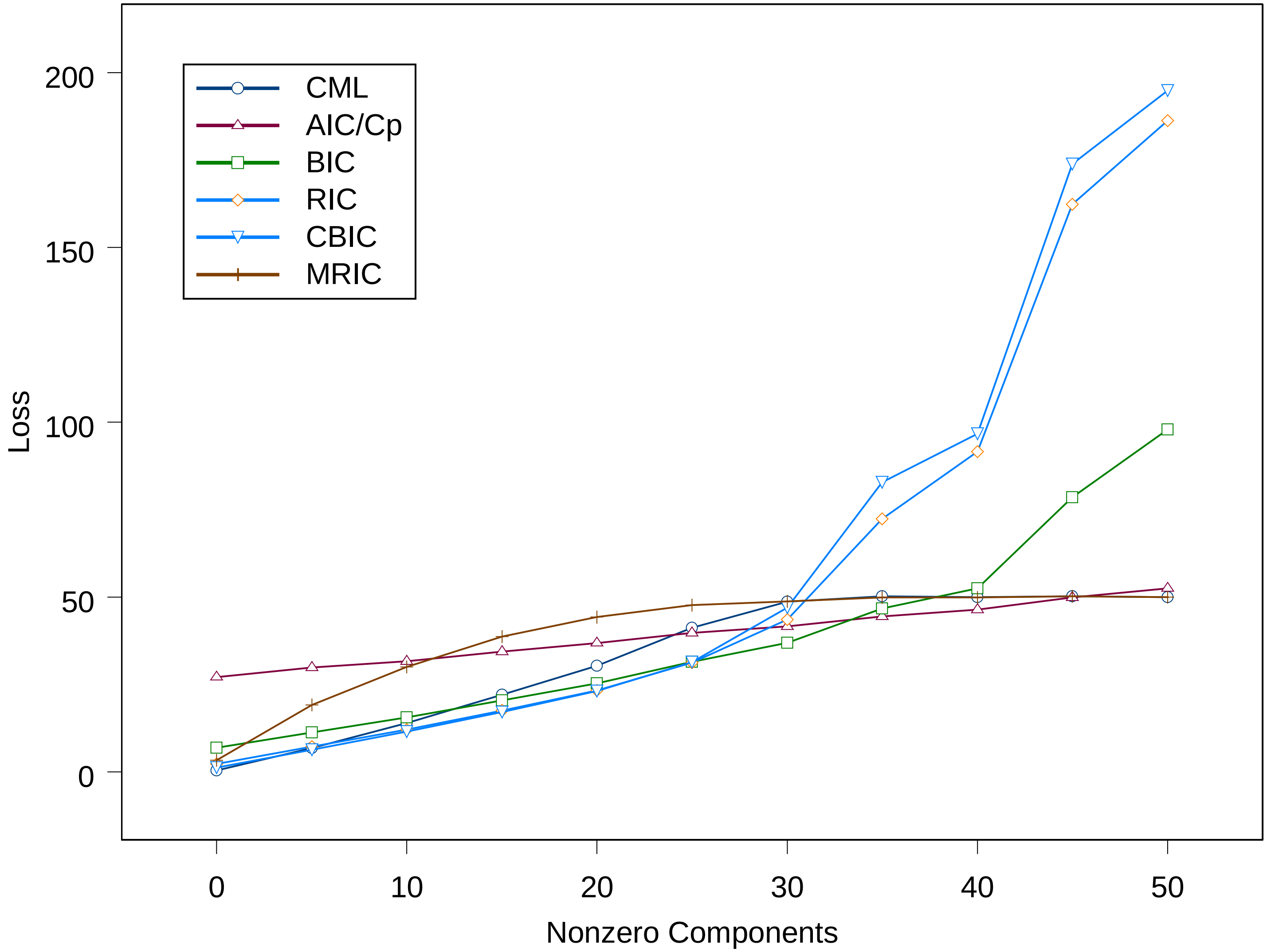

Figure 2(c). The average loss of the selection procedures when $\rho = -0.5$ for 11 choices of β with 0, 5, 10,...,50 nonzero components, respectively We denote C_{CML} by CML, Cauchy BIC by CBIC and modified RIC by MRIC.