International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2010), 78, 2, 240–256 doi:10.1111/j.1751-5823.2010.00108.x

On Model Selection Curves

Samuel Müller[1] and Alan H. Welsh[2]

[1]*School of Mathematics and Statistics F07, University of Sydney, NSW 2006, Australia*
E-mail: samuel.mueller@sydney.edu.au
[2]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia*

Summary

Many popular methods of model selection involve minimizing a penalized function of the data (such as the maximized log-likelihood or the residual sum of squares) over a set of models. The penalty in the criterion function is controlled by a penalty multiplier λ which determines the properties of the procedure. In this paper, we first review model selection criteria of the simple form "Loss + Penalty" and then propose studying such model selection criteria as functions of the penalty multiplier. This approach can be interpreted as exploring the stability of model selection criteria through what we call model selection curves. It leads to new insights into model selection and new proposals on how to select models. We use the bootstrap to enhance the basic model selection curve and develop convenient numerical and graphical summaries of the results. The methodology is illustrated on two data sets and supported by a small simulation. We show that the new methodology can outperform methods such as AIC and BIC which correspond to single points on a model selection curve.

Key words: Akaike Information Criterion (AIC); Bayesian Information Criterion (BIC); Generalized Information Criterion (GIC); linear regression; model selection; model selection curves.

1 Introduction

Suppose that we want to model the relationship between a response vector $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ matrix X using a linear regression model. Index the columns of X by $\{1, \ldots, p\}$ and then let α denote any subset of p_α distinct elements from $\{1, \ldots, p\}$. Let X_α denote the $n \times p_\alpha$ matrix with columns given by the columns of X whose indices appear in α and let $x_{\alpha i}^T$ denote the ith row of X_α. We assume the columnrank of X_α is p_α. Then, the linear regression model α is

$$y_i = x_{\alpha i}^T \beta_\alpha + \sigma_\alpha \epsilon_{\alpha i}, \quad i = 1, \ldots, n, \tag{1}$$

where β_α is an unknown p_α-vector of regression parameters, σ_α is an unknown spread parameter, X_α and $\epsilon_\alpha = (\epsilon_{\alpha 1}, \ldots, \epsilon_{\alpha n})^T$ are independent, and the $\epsilon_{\alpha i}$ are treated as having location zero and spread one. Let $\mathcal{A}$ denote a set of $N_\mathcal{A}$ statistical models for the relationship between y and X. (Both $\mathcal{A}$ and $N_\mathcal{A}$ can depend on n but we suppress this dependence for notational simplicity.) The purpose of model selection is to choose one or more models α from $\mathcal{A}$ with specified desirable properties.

Model selection is a fundamental problem in the practical application of statistics so there is an enormous and growing literature on the subject. We refer to Miller (2002) and Claeskens

& Hjort (2008) for a general introduction to subset selection in regression modelling and for an overview of model selection procedures. Many model selection strategies are based on the minimization over $\alpha \in \mathcal{A}$ of a loss function of the residuals plus a penalty term. The penalty term usually includes a penalty multiplier λ_n which controls the weight given to the penalty and determines the properties of the procedure.

In this paper we embed penalized loss functions into a more general framework of model selection curves that are functions of the penalty multiplier. Our approach is to analyse the whole curve rather than single points on the curve that is effectively the current strategy—see Section 2. The analysis of model selection curves gives us new insights into modelling and model selection, and allows us to select and order models subject to a particular property such as consistency or efficiency.

In Section 2, we briefly review penalized loss function criteria for model selection consisting of two additive terms, a description/prediction error component that we call the "Loss" term and a model complexity component that we call the "Penalty" term. In Section 3, we define what we mean by model selection curves. We show how to use the curves to capture the main aim of a model selection strategy, and then define three new model selection criteria. In Section 4, we introduce a framework for the selection of models using model selection curves. We show how to use the bootstrap to enhance the insights gained from model selection curves and discuss the presentation of the results. We then present the analysis of two real data sets and establish the usefulness of model selection curves. In Section 5, we report some simulation results, which demonstrate that analyzing model selection curves has the potential to outperform model selection criteria, which use a single value for the penalty multiplier. We conclude the paper with some brief remarks and conclusions in Section 6.

2 Review

In this paper, we consider model selection methods which choose models by minimizing an expression that can be written as "Loss + Penalty". In this Section, we discuss some of the many possible choices for both of these terms. Of course, there are other methods such as those based on adjusted R^2 which are not of this form and which we do not consider here.

The 'Loss'

The classical choice for the 'Loss' is minus twice the log-likelihood, $-2ll$. In the normal case, this leads to $n\log\{S_n(\alpha)/n\}$, where $S_n(\alpha) = \sum_{i=1}^{n}(\hat{y}_i(\alpha) - y_i)^2$ is the residual sum of squares and $\hat{y}(\alpha) = X_\alpha(X_\alpha^T X_\alpha)^{-1}X_\alpha^T y$ are the fitted values from model α. If we assume further that the variance $\sigma_\alpha^2 = \sigma^2$ is known and constant for all models containing the non-redundant variables, then, up to constant terms, minus twice the log-likelihood is $-2ll \propto S_n(\alpha)/\sigma^2$. This loss function is often applied when σ^2 is unknown by estimating σ^2 from a fixed, baseline model α_f. The fixed model typically has large columnrank p_{α_f} and is often chosen to be the full model $p_{\alpha_f} = p$ which contains all p columns of X. However, we do not insist that the fixed model be the full model because, when p is large compared to n, fitting the full model (and other very large models) may involve undesirable overfitting. Other "Loss" functions which can be used include log-quasilikelihoods, least squares, L_1 and other loss functions optimized in parameter estimation. Robust versions of these functions are of particular interest: see for example, Ronchetti & Staudte (1994), Konishi & Kitagawa (1996), Müller & Welsh (2005, 2009).

The 'Penalty'

The 'Penalty' term can penalize different aspects of the models. The simplest penalty is of the form $\lambda_n f_n(p_\alpha)$, where the penalty multiplier λ_n is a known, non-stochastic sequence and the penalty function f_n is a known, non-stochastic sequence of functions of the number of terms in the model p_α. For the simplest choice $f_n(p) = p$, we can choose the Akaike Information Criterion (AIC) penalty multiplier $\lambda_n = 2$ (Akaike, 1973; Mallows, 1973), the stepwise test penalty multiplier $\lambda_n = 4$ or, more generally, $\lambda_n \equiv \lambda$ (Shibata, 1984). The penalty multiplier can additionally depend on p_{α_f} (as in the Risk Inflation Criterion (RIC) penalty multiplier $\lambda_n = 2\log(p_{\alpha_f})$ of Foster & George, 1994), or on the sample size. The most common choices for the latter are $\lambda_n = c \log\log(n)$ (Hannan & Quinn, 1979), $\lambda_n/n \to 0$ and $\lambda_n/\log\log(n) \to \infty$ as $n \to \infty$ (Bai *et al.*, 1986; Rao & Wu, 1989); and the well-known Bayesian Information Criterion (BIC) penalty multiplier $\lambda_n = \log(n)$ (Schwarz, 1978) etc. There are a large number of other choices of $f_n(p)$: perhaps the best known of these is $f_n(p) = (p+1)/(n-p-2)$ used with $\lambda_n = 2$ by Sugiura (1978) and Hurvich & Tsai (1989).

Optimality

The penalties discussed earlier are derived under different optimization frameworks: For example, the AIC penalty $2p_\alpha$ minimizes the Kullback–Leibler distance between the model and the true density (Akaike, 1973) and is minimax optimal for estimating the regression function (Yang, 2005); the RIC penalty $2\log(p_{\alpha_f})p_\alpha$ asymptotically minimizes the maximum predictive risk inflation due to selection when the columns of X are orthogonal; and the BIC penalty $\log(n)p_\alpha$ optimizes the posterior probability of the model. Consistent selection of the true model, the model that actually generated the data, is also a useful criterion. The requisite asymptotic theory is well presented in Shao (1997) or Claeskens & Hjort (2008). If the true model is a linear regression model contained in the data we have observed (i.e. it is of the form (1)), we denote it by α_0. We usually require that $\alpha_0 \subseteq \alpha_f = \{1, \ldots, p_{\alpha_f}\}$ because this ensures that estimates produced by the fixed model are also valid estimates under the true model. If p_{α_0} is fixed, we get consistent model selection when $\lambda_n \to \infty$ such that $\lambda_n/n \to 0$ as $n \to \infty$ and we do not when $\lambda_n \to c$ for some constant c. On the other hand, if $p_{\alpha_0} \to \infty$ then the results are reversed. These results confirm the empirical experience that penalties with constant $\lambda_n = c$ (like the AIC penalty) produce methods which tend to select larger models and penalties with $\lambda_n \to \infty$ appropriately (like the BIC penalty) tend to choose smaller models, at least once n is large enough to ensure $\lambda_n > c$. In this sense, the optimal choice of penalty depends on the true model. Yang (2005) proved that we cannot find a λ_n which is able to achieve an adaptive compromise between the AIC penalty and the BIC penalty in the sense of achieving both consistency and minimax optimality. This result can be misinterpreted to imply that the AIC or BIC penalties are the only good choices for the penalty but this is not the case. As pointed out by Casella & Consonni (2009), Yang's result does not mean that the AIC or BIC penalties are the only good choices for the penalty when the power is not close to one and in general it is still well worth considering other choices.

Other 'Penalty's'

Other forms of penalty can also be considered, for example we can penalize the parameters themselves using $f_n(\beta_\alpha)$. The best known of these include the ridge penalty $f_n(\beta) = \sum_{j=1}^p \beta_j^2$, the Lasso penalty $f_n(\beta) = \sum_{j=1}^p |\beta_j|$ and the smoothing spline penalty. Other penalties include the Takeuchi Information Criterion (TIC) penalty (Takeuchi, 1976), the GIC penalty (Konishi & Kitagawa, 1996), the robust C_p penalty (Ronchetti & Staudte, 1994), the SIC penalty (Sugiyama

& Ogawa, 2001), etc. These penalties are generally related to the choice of 'Loss' and so tend to be used in rather specific criteria.

Relationship to Resampling Procedures

Cross validation and the bootstrap can be used either to measure the uncertainty of estimated best models obtained from other model selection procedures or as model selection procedures in their own right. Shao (1993, 1996, 1997) considered selection procedures using cross validation in which the model is fitted to $m \leq n$ observations and evaluated on the remaining $n - m$ observations and the m-out-of-n bootstrap in which the model is fitted to m observations sampled independently with replacement from the data and then evaluated on the original n observations. He showed that the asymptotic behaviour of these procedures is analogous to that of the penalized loss function procedures: leave-one-out cross validation ($m = n - 1$) and the $m = n$ bootstrap behave asymptotically like procedures with the AIC penalty (for leave one-out cross validation, see also Li, 1987) while cross validation and the m-out-of-n bootstrap with $m \to \infty, m/n \to 0$ as $n \to \infty$ behave asymptotically like procedures with the BIC penalty. Ronchetti *et al.* (1997) robustified the cross validation model selection procedure and Müller & Welsh (2005) presented a robust bootstrap model selection criterion. Müller & Welsh (2005) also proposed using stratified resampling to control the proportion of outliers in the bootstrap samples and ensure that these samples mimic the original data better. This is achieved by stratifying the data (so that observations with residuals in the tails of the residual distribution are placed in upper and lower tail strata and the remaining observations in other strata) and then resampling independently within each stratum (for more details see Section 3 of Müller & Welsh, 2005).

3 Model Selection Curves

The approach we now develop can be applied to general "Loss + Penalty" criteria but it is helpful, for definiteness, to fix attention on a particular criterion. We consider the generalized information criterion of Shao (1997) that is equivalent to

$$M_n(\lambda; \alpha) = \frac{S_n(\alpha)}{\hat{\sigma}_n^2} + \lambda p_\alpha, \qquad \lambda \geq 0, \alpha \in \mathcal{A}, \tag{2}$$

where $S_n(\alpha) = \sum_{i=1}^n (\hat{y}_i(\alpha) - y_i)^2$ and $\hat{\sigma}_n^2$ is an estimator of the spread parameter based on the residuals from the fixed model α_f. The generalized information criterion (GIC) is applied by choosing a specified function λ_n of n and then selecting the model $\hat{\alpha}(\lambda_n)$ that minimizes $M_n(\lambda_n; \alpha)$ over $\alpha \in \mathcal{A}$. In a slight abuse of terminology, we refer to GIC with $\lambda = 2$ as AIC and GIC with $\lambda = \log(n)$ as BIC.

From a practical point of view it is important for a model selection procedure to be stable or at least for a user to be aware when the procedure is unstable. There are various ways to define stability but a key idea is that small changes should have only small effects. Formally, we say that a model selection procedure is *unstable* when we select a model $\hat{\alpha}(\lambda_n)$ with dimension $p_{\hat{\alpha}(\lambda_n)}$ but, for some small $\delta > 0$, we select $\hat{\alpha}(\lambda_n + \delta)$ with smaller dimension $p_{\hat{\alpha}(\lambda_n+\delta)} < p_{\hat{\alpha}(\lambda_n)}$ and *stable* otherwise. To inform ourselves about stability, we investigate what happens to a model selection criterion in a neighborhood of λ_n by using model selection curves.

From the definition (2), for fixed α, $M_n(\lambda; \alpha)$ is a linear function of λ with intercept $M_n(0; \alpha) = S_n(\alpha)/\hat{\sigma}_n^2$ and slope p_α, both depending on α. Suppose that $M_n(0; \alpha)$ takes on $m \leq N_{\mathcal{A}}$ different values so we can reduce $\mathcal{A}$ to a set $\mathcal{A}_m^* = \{\alpha_1^*, \ldots, \alpha_m^*\}$ of the m smallest models that correspond to these m baseline values. (The largest model in $\mathcal{A}$ (which is often the fixed model α_f) appears as α_1^* when the GIC criterion (2) is used but not necessarily if we use other criteria.)

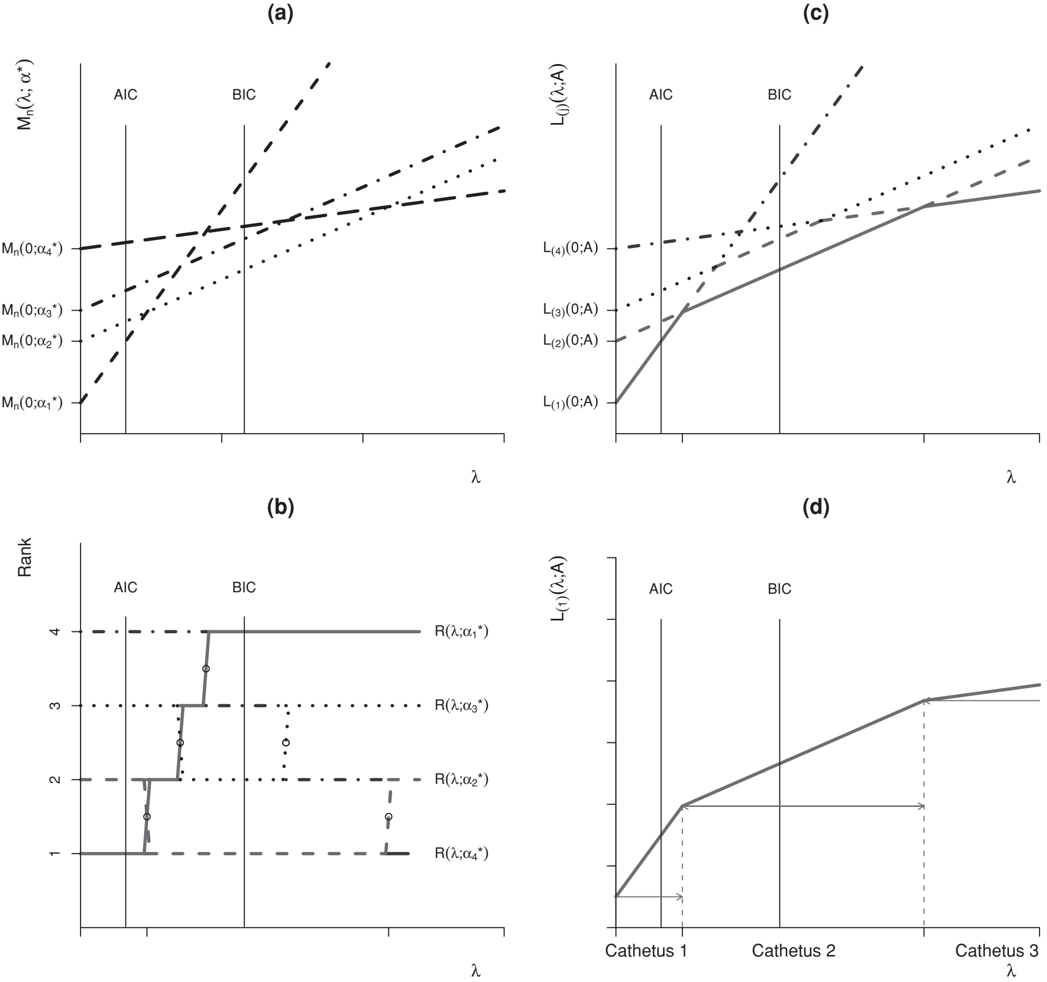

Figure 1. *An artificial example illustrating the construction of model selection curves. Panel (a) shows $m = 4$ curves $M_n(\lambda;\alpha)$ which correspond to models $\{\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*\}$; Panel (b) shows the corresponding rank functions for each model; Panel (c) shows the four rank selection curves $L_{(j)}(\lambda;\mathcal{A})$; Panel (d) shows the model selection curve and the definition of the three catheti. The values of λ for AIC and BIC are indicated by labelled vertical lines on the plots.*

Figure 1(a) shows an artificial example with $m = 4$. The four functions $M_n(\lambda;\alpha_j^*)$, $j = 1, \ldots, 4$ are plotted as straight lines against λ. The standard approach to model selection of fixing the penalty multiplier corresponds to drawing a vertical line through λ and then ordering or ranking the m models by the order in which the m lines $M_n(\lambda;\alpha_j^*)$ cross the vertical line. Thus, in Figure 1(a), for $\lambda = 2$ (corresponding to AIC), the models are ranked $(\alpha_1^*, \alpha_2^*, \alpha_3^*, \alpha_4^*)$ and GIC = AIC selects α_1^*, while for $\lambda = \log n$ (corresponding to BIC), the models are ranked $(\alpha_2^*, \alpha_3^*, \alpha_4^*, \alpha_1^*)$ and GIC = BIC selects α_2^*. We can carry out this ranking process for each λ to show how the rank of a model α changes as λ changes. Formally, for each $\lambda \geq 0$, let the *rank function* $R(\lambda;\alpha) = \text{rank}(M_n(\lambda;\alpha))$ be the rank of $M_n(\lambda;\alpha)$ for $\alpha \in \mathcal{A}_m^*$. The rank functions for each of the four models in the artificial example are shown in Figure 1(b). They are step functions, pairs of which have jumps at the values of λ at which the corresponding lines $M_n(\lambda;\alpha)$ cross.

We define *k rank model selection curves* $L_{(k)}(\lambda; \mathcal{A})$ for all λ where no ties of the m values $\{M_n(\lambda; \alpha_j^*)\}$ occur by

$$L_{(k)}(\lambda; \mathcal{A}) = \max\{M_n(\lambda; \alpha); \alpha \in \mathcal{A}_m \wedge R(\lambda; \alpha) \leq k\}, \tag{3}$$

and extend the definition to points at which ties occur simply by ensuring that $L_{(k)}(\lambda; \mathcal{A})$ is continuous over the entire range of $\lambda > 0$. The four rank selection curves for the artificial example are shown in Figure 1(c). The 1 rank model selection curve is the lower enveloping curve that can be defined equivalently as

$$L_{(1)}(\lambda; \mathcal{A}) = \min\{M_n(\lambda; \alpha); \alpha \in \mathcal{A}\}, \tag{4}$$

and we refer to it simply as the *model selection curve* for convenience. The model selection curve for the artificial example is shown in Figure 1(d). We can interpret $L_{(1)}(\lambda; \mathcal{A})$ as the entire solution path of the model selection criterion as a function of λ.

Remark 1. *Rank selection curves have some nice geometric properties. Here we mention just two.*

i. *The 1 rank selection curve $L_{(1)}(\lambda; \mathcal{A})$ is a convex polygon with at most $\#(p_\alpha) - 1$ knot points, where $\#(p_\alpha)$ is the number of distinct values of p_α for $\alpha \in \mathcal{A}$; by construction of the GIC criterion in (2) all $N_{\mathcal{A}}$ curves are lines with slope p_α, so the number of distinct values of p_α determines the maximum number of distinct slopes.*
ii. *The 2 rank selection curve $L_{(2)}(\lambda; \mathcal{A})$ is a piecewise convex polygon; it is bounded from below by $L_{(1)}$ so this is a direct consequence of the previous remark. In particular $L_{(2)}(\lambda; \mathcal{A})$ is a convex polygon on consecutive points of $L_{(2)}(\lambda; \mathcal{A}) \cap L_{(1)}(\lambda; \mathcal{A})$.*

Remark 1(i) implies that, regardless of the size of $\mathcal{A}$ or $\mathcal{A}_m^*$, at most $\#(p_\alpha) \leq p$ models (which we call the *candidate models*) can appear on the model selection curve. As the GIC function $M_n(\lambda; \alpha)$ has the same slope for all models with the same size p_α, the only model from each size class which can appear on the model selection curve is the model with smallest $M_n(0; \alpha)$. The candidate models can therefore be identified by computing $M_n(0; \alpha)$ and then ranking the models at $\lambda = 0$. Limiting the range of λ values (as we need to) further reduces the number of candidate models. Thus the linear structure of $M_n(\lambda; \alpha)$ greatly simplifies the computation of model selection curves. With more complicated criteria than (2), we may need to carry out the computations on a grid of λ values.

For each λ_n, GIC selects the model $\alpha \in \mathcal{A}_m^*$ for which $M_n(\lambda_n; \alpha) = L_{(1)}(\lambda_n; \mathcal{A})$, or equivalently for which $R(\lambda_n; \alpha) = 1$. This can be written in the form

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}}{\text{argmin}} \int \eta_\alpha(R(\lambda; \alpha))\, d\Delta_n, \tag{5}$$

where $\eta_\alpha(x) = 1 - \mathbf{1}\{x = 1\}$ and Δ_n is the Dirac measure that puts mass 1 on the point $\lambda = \lambda_n$. We can use the whole model selection curve rather than a single point on it by replacing the Dirac measure for the penalty multiplier λ in (5) by other measures such as the uniform distribution on $[\lambda_n^{\min}, \lambda_n^{\max}]$. Geometrically, this measure selects a model α that corresponds to the *longest cathetus* (LC) in the truncated polygon $L_{(1)}(\lambda; \mathcal{A})$, where the truncation is according to the interval $[\lambda_n^{\min}, \lambda_n^{\max}]$. (A cathetus is a side of a right angle triangle adjacent to the right angle: here the relevant cathetus is the horizontal edge of the right angle triangle whose hypotenuse is a segment of $L_{(1)}(\lambda; \mathcal{A})$. For the artificial example, the 3 catheti are shown in Figure 1(d).) The *longest edge* (LE) in the truncated polygon (i.e. the length of the hypotenuse in the right angle triangles from which the catheti are extracted) corresponds to a philosophy that leans towards larger models than models selected by the LC criterion, that is the LC criterion is more parsimonious than LE. The model selected by the longest edge criterion is

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathcal{A}} \int \sqrt{1 + p_{\alpha_f}^2} - \sqrt{1 + p_{\alpha}^2} \cdot \mathbf{1}\{R(\lambda; \alpha) = 1\}\, d\lambda. \tag{6}$$

A somewhat different selection procedure that is sequential and adaptive in nature is the *first substantially long cathetus* (FSLC) method, which selects the largest model that corresponds to an edge of the truncated polygon $L_{(1)}(\lambda; \mathcal{A})$ with minimal cathetus length of say $(c/4)\log(n),\ c \in \mathbb{N}$, if one exists, and selects the model selected by the LC criterion otherwise.

Under a mild condition on the penalty measure used to define the LC and longest edge criteria, we can show that they are consistent.

LEMMA 1. *Let p_{α_0} be fixed, assume $\xi_n < \lambda_n^{\max}$ satisfy $\xi_n \to \infty$ and $\lambda_n^{\max}/n \to 0$ as $n \to \infty$, and let the penalty measure satisfy $\Lambda_n([\lambda_n^{\min} + \xi_n, \lambda_n^{\max}]) \to 1$. Then the LC and LE selection criteria are consistent for α_0.*

Proof. For p_{α_0} fixed, GIC is consistent when $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ as $n \to \infty$ so it is consistent for $\lambda_n = \lambda_n^{\min} + \xi_n$ and for $\lambda_n = \lambda_n^{\max}$. Thus,

$$\int_{\lambda_n^{\min}}^{\lambda_n^{\max}} (1 - \mathbf{1}\{R(\lambda; \alpha_0) = 1\})\, d\Lambda_n \leq \int_{\lambda_n^{\min}}^{\lambda_n^{\min}+\xi_n} 1 d\Lambda_n + \int_{\lambda_n^{\min}+\xi_n}^{\lambda_n^{\max}} (1 - \mathbf{1}\{R(\lambda; \alpha_0) = 1\})\, d\Lambda_n.$$

The first term is $o_p(1)$ and the second is bounded from above by

$$\int_{\lambda_n^{\min}+\xi_n}^{\lambda_n^{\max}} (1 - \mathbf{1}\{R(\lambda_n^{\min} + \xi_n; \alpha_0) = 1\})\, d\Lambda_n + \int_{\lambda_n^{\min}+\xi_n}^{\lambda_n^{\max}} (1 - \mathbf{1}\{R(\lambda_n^{\max}; \alpha_0) = 1\})\, d\Lambda_n,$$

which is $o_p(1)$. Similarly, for models $\alpha \neq \alpha_0$, $\int_{\lambda_n^{\min}}^{\lambda_n^{\max}} (1 - \mathbf{1}\{R(\lambda; \alpha_0) = 1\}) d\Lambda_n = 1 + o_p(1)$. It follows that LC is consistent and this implies the consistency of LE. □

The lemma applies for example, when Λ_n is the Lebesgue measure, $\lambda_n^{\min} = 0,\ \xi_n = \log\log n$ and $\lambda_n^{\max} = 4\log(n)$. The result follows from the fact that GIC with the Hannan–Quinn penalty multiplier ξ_n and GIC with the (scaled) BIC penalty multiplier $\lambda_n^{\max}$ are consistent and $\log\log n/\log n = o(1)$.

We can introduce a wide range of new model selection criteria based on $L_{(1)}(\lambda; \mathcal{A}), \ldots, L_{(m)}(\lambda; \mathcal{A}),\ R(\lambda; \alpha)$ or $R^{-1}(1; \alpha)$, which gives the values of λ for which α is the selected model. In addition, we may be able to gain insight by comparing these criteria based on different measures of description, such as other more robust measures. We do not pursue these possibilities here.

As with any model selection criterion, if s columns of X are part of any model that makes sense, then those s columns can be forced to be part of all models under consideration in $\mathcal{A}$. In this case, we may reduce all penalties by multiplying λ_n by the number of free parameters $p_\alpha - s$ instead of p_α, making the true model easier to detect.

4 Enhancing and Using Model Selection Curves

In this section we discuss the use of the bootstrap to enhance model selection curves $L_{(1)}(\lambda; \mathcal{A})$, useful ways of presenting the results and two examples illustrating the use of the methodology.

The model selection curve $L_{(1)}(\lambda; \mathcal{A})$ is usefully summarized by tabulating the catheti and their absolute and relative length on the interval $[\lambda_n^{\min} = 0, \lambda_n^{\max} = 4\log(n)]$. The upper truncation point is arbitrary but is chosen to contain the common choices of λ_n shown in Section 2. For a cathetus, we denote the x-coordinate of the lower endpoint l_α and the x-coordinate of the upper endpoint u_α so that the length of the cathetus is $\delta_\alpha = u_\alpha - l_\alpha$ and the relative length is

$$\omega_\alpha = \frac{\delta_\alpha}{\lambda_n^{\max} - \lambda_n^{\min}} = \frac{\delta_\alpha}{4\log(n)}. \tag{7}$$

Both δ_α and ω_α measure the range of values of λ over which the model selection procedure selects the same model and hence present evidence that the model generating this cathetus is the most stable choice of model.

We use the bootstrap to estimate model detectability as a function of the penalty multiplier by estimating the probability that model $\alpha \in \mathcal{A}$ corresponds to the point on the model selection curve at λ. We do this by calculating the model selection curve $L_{(1)}(\lambda; \mathcal{A})$ for stratified bootstrap samples from the observed data. We prefer using stratification based on the residuals from the fixed model α_f because it preserves the robustness properties of estimators (see e.g. Müller & Welsh, 2005, 2009). The model selection curve for a bootstrap sample can include models which were not represented on the model selection curve for the observed data so bootstrapping can expand the number of candidate models. This occurs because the ranking at $\lambda = 0$ of the GIC criterion function (2) can change in different bootstrap samples, allowing additional models to appear on bootstrap model selection curves. In our experience, this expansion is small and the total number of candidate models is usually very small, as illustrated in our simulation study in Section 5.

The empirical bootstrap estimate $\pi^*(\lambda; \alpha)$ of the probability of selecting model α when we use the penalty multiplier λ can be used to compute $\pi^*(\alpha) = \int_{\lambda_n^{\min}}^{\lambda_n^{\max}} \pi^*(\lambda; \alpha)\, d\Lambda_n$, where for simplicity Λ_n is $\mathcal{U}(\lambda_n^{\min}, \lambda_n^{\max})$. Here $\pi^*(\alpha)$ is a bootstrap estimate of the marginal probability of selecting model α. A comparatively large ω_α does not imply a large $\pi^*(\alpha)$ and $\omega_\alpha = 0$ (i.e. model α does not produce an edge on the model selection curve) does not imply that model α is not a possible optimal model, unless $\pi^*(\alpha)$ is also close to zero. The bootstrap probability $\pi^*(\alpha)$ is somewhat related to the coverage probability of a confidence interval in classical statistical estimation as it approximates the probability that a given model $\alpha \in \mathcal{A}$ generates an edge on the model selection curve. As not all models that have large $\pi^*(\alpha)$ are part of the model selection curve, we consider all models with $\pi^*(\alpha) > 4\%$, allowing at most $0.04^{-1} - 1 = 24$ additional models that do not appear on the model selection curve, although based on our experience it is very unlikely that more than three to five additional models have to be taken into account. (It is tempting to interpret $\pi^*(\alpha) < q$ as the critical region of a level q test of the hypothesis that α is the best model in $\mathcal{A}$ but this is not true in a mathematical sense.) If a model has clearly the largest $\pi^*(\alpha)$, particularly if it is greater than 50%, then this is a strong indication that for $\lambda \in [\lambda_n^{\min}, \lambda_n^{\max}]$, this is the best model.

Aggregating $\pi^*(\lambda; \alpha)$ or $\pi^*(\alpha)$ over models with the same dimension or over models that have at least one common variable gives access to diagnostic measurements for the best model dimensionality and for statements regarding the inclusion or exclusion of particular variables in the final model. More precisely by aggregating over dimension $j = 1, \ldots, p_{\alpha_f}$ we mean calculating

$$\pi_j^{*d}(\lambda) = \sum_{\alpha: p_\alpha = j} \pi^*(\lambda; \alpha) \quad \text{or} \quad \pi_j^{*d} = \sum_{\alpha: p_\alpha = j} \pi^*(\alpha),$$

and by aggregating over a particular variable $k = 1, \ldots, p_{\alpha_f}$ we mean calculating

$$\pi_k^{*v}(\lambda) = \sum_{\alpha: \alpha \ni k} \pi^*(\lambda; \alpha) \quad \text{or} \quad \pi_k^{*v} = \sum_{\alpha: \alpha \ni k} \pi^*(\alpha).$$

There are many ways of visualizing information from a model selection curve. We have found the following four plots of particular value for (1) determining a single best or a few candidates for the best model, (2) determining the optimal model dimension, and (3) ranking and quantifying the importance and contribution of variables to the final list of best model candidates.

(a) Rank plot: $R(\lambda; \alpha)$ against λ for the candidate models α.
(b) Single model plot: $\pi^*(\lambda; \alpha)$ against λ for the candidate models α and other models which have $\bar{\pi}^*_\alpha > 4\%$.
(c) Dimensionality plot: $\pi^{*d}_j(\lambda)$ against λ for all j.
(d) Variable detection plot: $\pi^{*v}_k(\lambda)$ against λ for all k.

The variable detection plot is like the Lasso coefficient path plot (see e.g. Hastie *et al.*, 2001). We illustrate the use of these plots in two examples.

4.1 Cement Data

We first consider the well-known data of Woods *et al.* (1932) on the heat generated during the hardening of Portland cement. Following Hald (1952; pp. 635–649), the heat generated is traditionally modelled as a function of four chemicals; we refer to Piepel & Redgate (1998) for an analysis based on the original data that uses more than four explanatory variables. We use the same $n = 13$ observation vectors as shown in Table 5.1 in Flury & Riedwyl (1988) with components heat evolved (Y), tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3), and dicalcium silicate (x_4). For simplicity, we denote the intercept vector as x_0 and, as we include the intercept in all our models, omit it when we describe a model. Thus $\{2, 4\}$ is shorthand for $\{0, 2, 4\}$ etc. With $\lambda_n^{\max} = 4 \log(13)$, three models appear on the model selection curve, $\alpha_1^* = \alpha_f = \{1, 2, 3, 4\}$, $\alpha_2^* = \{1, 2, 4\}$ and $\alpha_3^* = \{1, 2\}$. Three further models appear frequently, i.e. they have $\pi^*(\alpha) > 0.04$. There are only 16 possible models with an intercept so we include all models with $\pi^*(\alpha) > 0.01$ in the diagnostic Table 1. The observations were assigned to three strata according to the rank of their absolute residual from the fixed model $\{1, 2, 3, 4\}$: the first five central residuals were placed in the first stratum and the four intermediate and the four most extreme residuals were placed in strata two and three, respectively. We carried out 1,000 bootstrap replications.

Based on Table 1, the best model is $\{1, 2\}$ which is also the model selected by AIC and BIC. It is interesting to note that the full model has little support. The model $\{1, 2, 4\}$ is the least likely among the models with three slope parameters but it appears on the model selection curve while $\{1, 2, 3\}$ does not. According to the bootstrap, $\{1, 2, 3\}$ is five times more likely than $\{1, 2, 4\}$. We conclude that a model with dimension four is not as good as one with dimension three.

Table 1
Diagnostic table for Hald data.

Model α	p_α	l_α	u_α	δ_α	ω_α	$\pi^*(\alpha)$
$\{1, 2, 3, 4\}$	5	0	0.01	0.01	0.00	0.05
$\{2, 3, 4\}$	4	-	-	-	-	0.04
$\{1, 3, 4\}$	4	-	-	-	-	0.03
$\{1, 2, 4\}$	4	0.02	1.66	1.64	0.16	0.02
$\{1, 2, 3\}$	4	-	-	-	-	0.11
$\{2, 4\}$	3	-	-	-	-	0.02
$\{1, 4\}$	3	-	-	-	-	0.21
$\{1, 2\}$	3	1.67	10.26	8.59	0.84	0.53

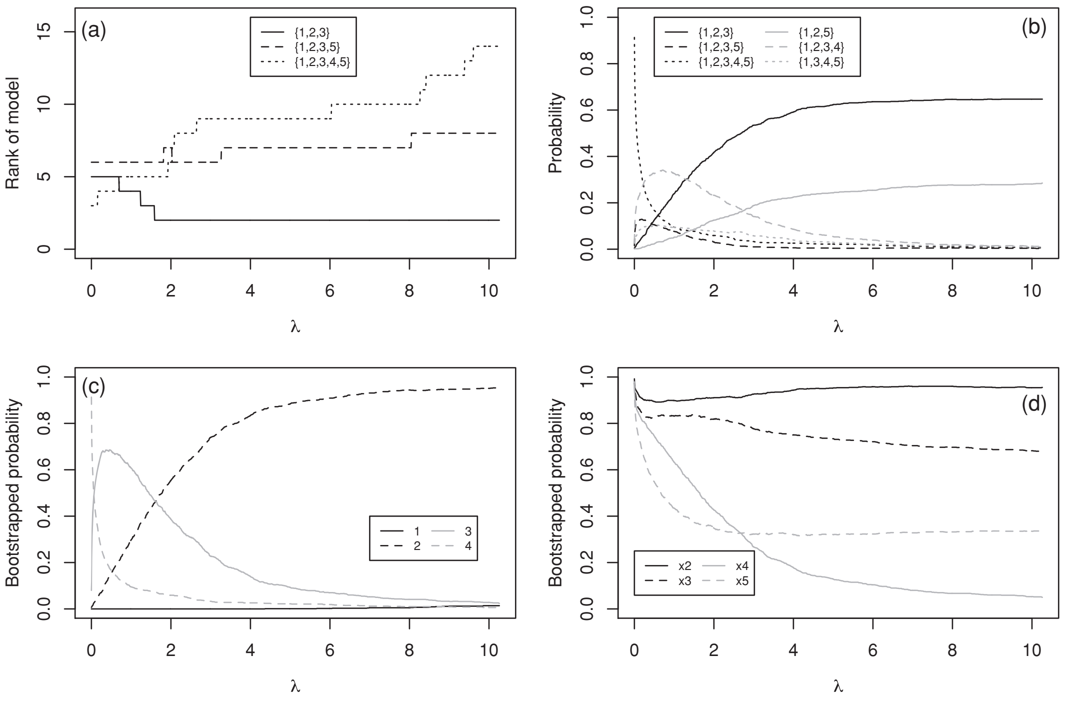

Figure 2. *Diagnostic plots for cement data: (a) rank plot; (b) single model plot; (c) dimensionality plot; (d) variable detection plot.*

This is supported by the aggregated bootstrap probabilities by dimension which provide clear evidence that the best model should have dimension 3 (i.e. two slope parameters):

$$\pi_5^{*d} = 0.05, \quad \pi_4^{*d} = 0.19, \quad \pi_3^{*d} = 0.75, \quad \pi_2^{*d} = 0.00, \quad \pi_1^{*d} = 0.00.$$

The aggregated bootstrap probabilities by variable reveal that x_1 and x_2 are clearly much more important than x_3 and x_4, confirming the chosen best model:

$$\pi_1^{*v} = 0.94, \quad \pi_2^{*v} = 0.75, \quad \pi_3^{*v} = 0.24, \quad \pi_4^{*v} = 0.36.$$

The four diagnostic plots in Figure 2 support these findings. We make several comments about the plots.

(a) Rank plot (top left): Plots the rank function $R(\lambda; \alpha)$ against λ for models α on the model selection curve and shows the catheti. The figure also shows that, for the full model, $R(\lambda; \{1, 2, 3, 4\})$ is almost linearly increasing in λ, that for model $\{1, 2\}$, $R(\lambda; \{1, 2\})$ linearly decreases until it plateaus at minimal rank for $\lambda \geq 1.67$. Model $\{1, 2, 4\}$ almost behaves as the full model is expected to behave, that is to have minimal rank on $[0, c]$, $0 < c < 2$ and then to increase until all smaller models have smaller rank.

(b) Single model plot (top right): Plots the bootstrap probability $\pi^*(\lambda; \alpha)$ against λ for models α on the model selection curve or with $\pi^*(\alpha) > 4\%$. It is interesting to note that model $\{1, 2\}$ has largest $\pi^*(\lambda; \alpha)$ slightly before it appears on the model selection curve and is considerably larger on the entire interval $[1.67, \lambda_n^{\max}]$. Out of the models that appear on the model selection curve, $\{1, 2\}$ has the largest bootstrap probability for λ's as small as around 0.7.

(c) Dimensionality plot (bottom left): Plots $\pi_j^{*d}(\lambda)$ against λ for all j. The dimensionality plot reveals that a model with dimension 2 is the natural dimension if the main purpose is to choose a model with good predictive ability. On the other hand a model of dimension three could be of some additional value if the main focus is on the description of the data. We base this statement on the fact that for $\lambda < 2$ a model of dimension three is more or approximately as likely as a model of dimension two.

(d) Variable detection plot (bottom right): Plots $\pi_k^{*v}(\lambda)$ against λ for all k. It is interesting that x_4 becomes more important than x_3 for large λ suggesting that it is more important for prediction but x_3 is better for description. This is underlined by the fact that $\pi^*(\{1, 2, 3\}) = 11\%$ even though this model does not appear on the model selection curve. The plot demonstrates clearly that both variables, x_1 and x_2 have excellent description and prediction qualities because their bootstrapped probabilities are almost constantly close to 1 and 0.8, respectively, over the full range of relevant λ's.

4.2 Physical Measurements

In order to demonstrate the usefulness of our diagnostic tools when the number of possible models is large, say greater than 1,000, we analyze the physical measurement data from the Australasian Data and Story Library. The data together with a brief description are available from `http://www.statsci.org/data/oz/physical.html`. The weight (Y) of $n = 22$ male subjects is potentially explained by 10 different physical measurements such as the maximum circumference of the forearm (x_1) or the maximum circumference of the head (x_{10}). As in the cement data example, we denote the intercept vector by x_0 and omit the intercept when specifying models. We consider all $2^{10} = 1,024$ possible linear regression models.

With $\lambda_n^{\max} = 4\log(22)$, nine models appear on the model selection curve but five of these are supported by catheti with length smaller than 1. Only five models appear frequently on the bootstrapped model selection curves with $\bar{\pi}_\alpha^* > 0.04$: two of these models are not on the model selection curve for the observed data. Model $\{1, 6\}$ is selected most frequently in the stratified bootstrap data with an empirical selection probability of 9.0%. This low value means that the determination of a single best model for this data set is, to say the least, challenging and more likely not sensible. In Table 2 we present the results for the 11 candidate models, an initial reduction of models by almost a factor of 100. The observations were assigned to four strata according to the rank of their absolute residual for the fixed model $\{1, \ldots, 10\}$: the first six and second six central residuals were assigned to two strata of size 6 each and the 10 most extreme residuals were assigned to two further strata of size 5 each. The number of bootstrap replications was again 1,000.

Table 2
Diagnostic table for physical measurement data.

Model α	p_α	l_α	u_α	δ_α	ω_α	$\pi^*(\alpha)$
$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$	11	0	0.01	0.01	0.001	0.030
$\{1, 2, 3, 4, 6, 7, 8, 9, 10\}$	10	0.02	0.11	0.09	0.009	0.002
$\{1, 3, 4, 6, 7, 8, 9, 10\}$	9	0.12	0.34	0.22	0.018	0.001
$\{1, 3, 6, 7, 8, 9, 10\}$	8	0.35	0.90	0.55	0.044	0.001
$\{1, 6, 7, 8, 9, 10\}$	7	0.91	1.01	0.10	0.001	0.005
$\{1, 6, 7, 9, 10\}$	6	1.02	2.28	1.26	0.102	0.016
$\{1, 6, 7, 9\}$	5	2.29	5.21	2.92	0.236	0.048
$\{4, 6, 8\}$	4	-	-	-	-	0.045
$\{1, 6\}$	3	5.22	11.32	6.10	0.494	0.090
$\{5, 6\}$	3	-	-	-	-	0.071
$\{1\}$	2	11.33	12.36	1.03	0.0.083	0.044

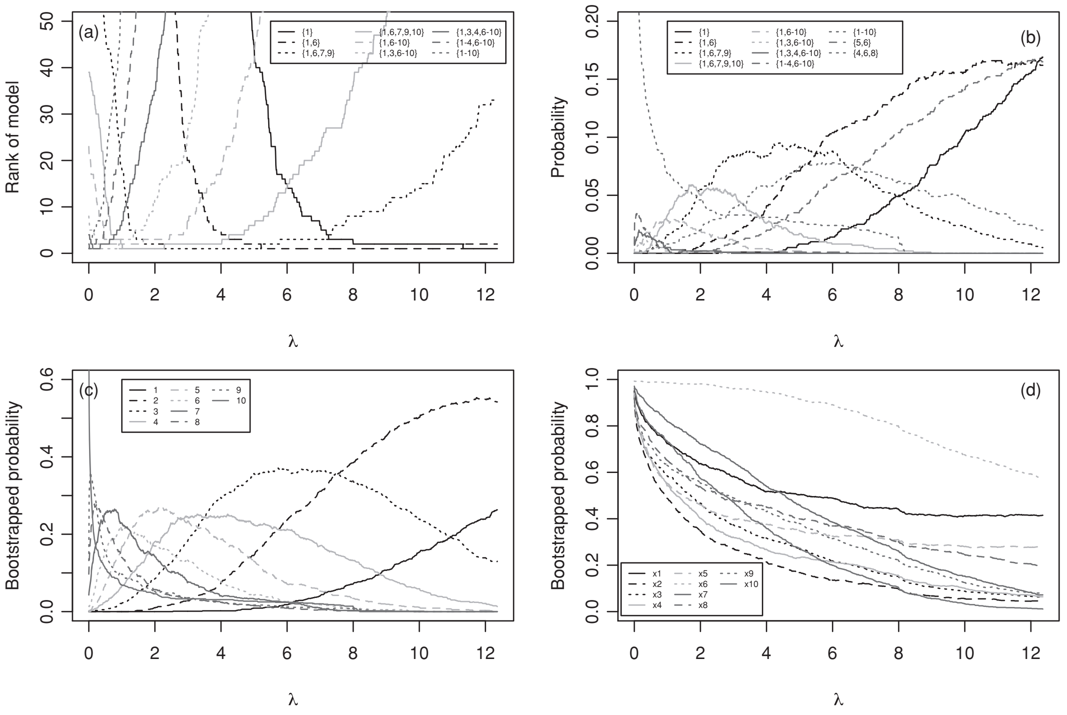

Figure 3. *Diagnostic plots for physical measurement data: (a) rank plot; (b) single model plot; (c) dimensionality plot; (d) variable detection plot.*

From Table 2, the best candidate models have dimension between 6 and 3 that enables a further reduction to five candidate models only. This is supported by the aggregated bootstrap probabilities by dimension which provide further evidence that the best model should have dimension 3–5 (i.e. 2–4 slope parameters):

$$\pi_{11}^{*d} = 0.030, \quad \pi_{10}^{*d} = 0.032, \quad \pi_{9}^{*d} = 0.034, \quad \pi_{8}^{*d} = 0.053, \quad \pi_{7}^{*d} = 0.059,$$
$$\pi_{6}^{*d} = 0.096, \quad \pi_{5}^{*d} = 0.135, \quad \pi_{4}^{*d} = 0.230, \quad \pi_{3}^{*d} = 0.256, \quad \pi_{2}^{*d} = 0.065.$$

The aggregated bootstrap probabilities by variable reveal that x_6 has the strongest claim ($\pi_6^{*v} = 83.9\%$) to be included in the selected model, followed by x_1 with $\pi_1^{*v} = 51.6\%$:

$$\pi_{1}^{*v} = 0.516, \quad \pi_{2}^{*} = 0.195, \quad \pi_{3}^{*v} = 0.265, \quad \pi_{4}^{*v} = 0.249, \quad \pi_{5}^{*v} = 0.370,$$
$$\pi_{6}^{*v} = 0.839, \quad \pi_{7}^{*v} = 0.418, \quad \pi_{8}^{*v} = 0.394, \quad \pi_{9}^{*v} = 0.336, \quad \pi_{10}^{*v} = 0.277.$$

The four diagnostic plots in Figure 3 support these findings and provide deeper insight.

(a) Rank plot (top left): $R(\lambda; \alpha)$ against λ for models α on the model selection curve. When there is a large number of models, the plot can be difficult to read. The full model has a rank of more than 50 for λ around 1 and consequently quickly *disappears* from the relevant section of this plot. The same is true for the second and third largest models. The fourth largest model stays a little longer in the relevant range having a rank around 10 for the AIC λ value of 2 and a rank around 20 for the BIC λ value of $\log(n) = 3.091$. The fifth largest model shows similar behaviour. This provides evidence that models of dimension larger than seven and possibly even six are clearly too large. The model chosen by the BIC $\{1, 6, 7, 9\}$

has low rank for λ as small as 1.5 and up to $\lambda = 2.5\log(n)$. The model of dimension two is not even in the top fifty for the AIC λ and enters the picture only later with low rank after the BIC λ value.

(b) Single model plot (top right): $\pi^*(\lambda;\alpha)$ against λ for models α on the model selection curve or with $\pi^*(\alpha) > 4\%$. Assuming that a model of dimension two is too small, we can neglect the additional model $\{5, 6\}$ and see that there are really two strong contenders, namely the model of dimension four $\{1, 6, 7, 9\}$ and the model of dimension three $\{4, 6, 8\}$.

(c) Dimensionality plot (bottom left): $\pi_j^{*d}(\lambda)$ against λ for all j. We see that models of dimension three to five dominate the most important range of λ, i.e. from the AIC value to about twice the BIC value. A model of dimension three is too small considering that the best model, $\{4, 6, 8\}$, is not well-embedded into the larger models as it has only x_6 in common with the best model of dimension four. Therefore we suggest restricting the natural dimension of the best model to dimension four or five.

(d) Variable detection plot (bottom right): $\pi_k^{*v}(\lambda)$ against λ for all k. Variable x_6 is clearly the dominating variable. Variable x_7 is very important for description but less important than x_1 for prediction. The variables x_5 and x_1 as well as x_8 and x_9 seem to be in competition with each other, but the latter pair seems to be better if x_6 is included in the model. The variables x_2, x_3, and x_4 seem to be the least important variables as can be seen from the aggregated bootstrap probabilities by variable. Variable x_{10} is excellent for description but least important for prediction.

5 Simulations

For comparison with other published results, we ran a simulation study based on the solid waste data of Gunst & Mason (1980). The same settings were used in Shao (1993, 1996, 1997), Wisnowski *et al.* (2003), and Müller & Welsh (2005) in the context of model selection. Thus, we consider the model

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i,$$

where $i = 1, \ldots, 40$, the errors ϵ_i are independent and identically distributed standard normal random variables, x_0 is the column of ones, and the values for the solid waste data variables x_1, x_2, x_3, and x_4 are taken from Shao (1993, Table 1). These explanatory variables are highly correlated as can be seen from their estimated Pearson correlation matrix

$$\widehat{\text{corr}}(X_{\{1,\ldots,4\}}) = \begin{pmatrix} 1.00 & 0.91 & 0.93 & 0.89 \\ 0.91 & 1.00 & 0.92 & 0.79 \\ 0.93 & 0.92 & 1.00 & 0.90 \\ 0.89 & 0.79 & 0.90 & 1.00 \end{pmatrix}.$$

We considered models with regression parameter $\beta = (2, 9, 6, 4, 8)^T$, $\beta = (2, 9, 0, 4, 8)^T$, $\beta = (2, 0, 0, 4, 8)^T$ and $\beta = (2, 0, 0, 4, 0)^T$. Hence the dimension of the true model varies between 5 and 2. In each of the 4×250 Monte Carlo simulation runs, we calculated the diagnostic table for the model selection curve, the relative length ω_α, the empirical bootstrap probability $\pi^*(\alpha)$ for all models with $\pi^*(\alpha) > 4\%$, and the aggregated bootstrap probability over model dimensionality π_j^{*d}. The number of bootstrap replications is again chosen to be 1,000. The observations were assigned to four strata of size 10 according to the rank of their absolute residual for the fixed model $\{1, 2, 3, 4\}$. With our R code the calculation of one simulation run takes of the order of 2–4 minutes (iMac, 2 GHz Intel Core 2 Duo, 2 GB 667 RAM): the time is influenced mostly by the number of models $N_{\mathcal{A}}$, the number of evaluated λ's, and the number of bootstrap

Table 3
Absolute and relative frequencies based on 250 simulation runs: (a) the true model appears on the model selection curve $M_{(1)}$; (b) the bootstrap probability for α_0 satisfies $\pi^*(\alpha_0) > 4\%$.

	(a) $\alpha_0 \in M_{(1)}$		(b) $\pi^*(\alpha_0) > 4\%$.	
True Model α_0 (Detectability)	abs	rel	abs	rel
$\{1, 2, 3, 4\}$	250	1.000	250	1.000
$\{1, 3, 4\}$	247	0.988	250	1.000
$\{3, 4\}$	250	1.000	250	1.000
$\{3\}$	250	1.000	250	1.000
Total	1,000	0.997	1,000	1.000

Table 4
Absolute and relative selection frequencies of the true model for some selection criteria: (a) AIC; (b) BIC; (c) longest cathedus; (d) maximal bootstrap probability $\max \pi^*(\alpha)$.

True Model α_0	(a) AIC		(b) BIC		(c) LC		(d) $\max \pi^*(\alpha)$	
was selected	abs	rel	abs	rel	abs	rel	abs	rel
$\{1, 2, 3, 4\}$	249	0.996	247	0.988	232	0.928	245	0.980
$\{1, 3, 4\}$	221	0.884	235	0.940	223	0.892	234	0.936
$\{3, 4\}$	175	0.700	225	0.900	248	0.992	235	0.940
$\{3\}$	153	0.612	214	0.856	246	0.984	236	0.944
Total	798	0.798	921	0.921	949	0.949	950	0.950

replications. We can only report some condensed information from this simulation study here, but all 1,000 diagnostic tables and the R code used in our simulation are available from the authors.

The first interesting result is that, in some of the simulation runs, the true model does not produce an edge on the model selection curve but is still identified as an important model by satisfying $\pi^*(\alpha_0) > 4\%$. With the least squares estimator, this can only happen for true models smaller than the fixed model.

In Table 3, we report the frequency with which selected models were detected. In all 1,000 simulation runs, the true model has $\pi^*(\alpha_0) > 4\%$. In three simulation runs the true model $\{1, 3, 4\}$ does not appear on the model selection curve: For runs 14, 69, and 150, the $\pi^*(\alpha_0)$ values are 5.06%, 28.12%, and 11.4%, respectively. This means that a model other than the true model can have the largest $\pi^*(\alpha)$ for some realizations.

We report in Table 4 the frequency with which selected models were detected by AIC, BIC, the LC criterion and the criterion based on maximizing $\pi^*(\alpha)$.

The simulation shows that using $\pi^*(\alpha)$ as a single criterion has the potential to outperform fixed λ_n GIC criteria even for standard situations (normal errors, least squares estimators). The LC criterion performs particularly well if the number of parameters in the true model are half or less than the number of parameters in the full model. As the frequencies in Table 4 suggest, there are a number of simulation runs where, particularly for the true model $\{3\}$, neither AIC nor BIC select the true model but both LC and $\pi^*(\alpha)$ select the correct model.

Table 5
Diagnostic table for simulation run 1 for true model $\{3, 4\}$.

Model α	p_α	l_α	u_α	δ_α	ω_α	$\pi^*(\alpha)$
$\{1, 2, 3, 4\}$	5	0	0.16	0.16	0.01	0.019
$\{2, 3, 4\}$	4	0.16	3.23	3.07	0.21	0.255
$\{3, 4\}$	3	3.23	14.76	11.53	0.78	0.673

Table 6
Diagnostic table for simulation run 12 for true model {3}.

Model α	p_α	l_α	u_α	δ_α	ω_α	$\pi^*(\alpha)$
$\{1, 2, 3, 4\}$	5	0	0.00	0.00	0.00	0.050
$\{1, 2, 3\}$	4	0.00	1.96	1.96	0.13	0.180
$\{2, 3\}$	3	1.96	4.44	2.48	0.17	0.133
$\{3, 4\}$	3	-	-	-	-	0.041
$\{3\}$	2	4.44	14.76	10.32	0.70	0.519

We report for illustration the diagnostic tables for two typical simulations runs, namely run 1 for the true model {3, 4} (Table 5) and run 12 for the true model {3} (Table 6). In Table 5, there were no additional models with $\pi^*(\alpha) > 4\%$ whereas, in Table 6, the four models on the model selection curve were supplemented by an additional model, namely {3, 4} with $\pi^*(\{3, 4\})$ just over 4%.

The conclusion from this simulation study is that, in most runs, the true model is clearly chosen as the best model in that $\pi^*(\alpha_0)$ is larger than 50% and all the other bootstrap probabilities are small. This outcome is reflected in the model having a large corresponding edge on the model selection curve. A small fraction of the simulation runs show inconclusive diagnostic tables. In those cases there are typically two or at most three remaining candidate models that can be analyzed further.

6 Conclusions

In this paper, we have embedded model selection criteria used in methods like AIC and BIC inside a model selection curve which enables us to study the criterion function as a function of the penalty multiplier instead of simply at single values of the penalty multiplier. This approach allows us to explore the stability of criterion functions and hence selected models. It leads to new insights into model selection and new proposals on how to use model selection curves to select models. We used the bootstrap to enhance the basic model selection curve and developed convenient numerical and graphical summaries of the results. We illustrated the methodology on two data sets and in a small simulation study.

We argue that model selection curves are both philosophically and practically important. In the first case, when we select a particular point on a model selection curve (as when we use AIC, BIC, etc.), we obtain a solution that is explicitly or implicitly linked to a specific point of view on predictive versus descriptive performance. Using the whole curve is less tied to a specific viewpoint and, in fact, when enhanced by bootstrapping, gives useful insight into these tradeoffs. In the second case, using the whole curve with enhancement can outperform single point methods. Simulation examples can be constructed to show the superiority of model selection criteria based on the model selection curve over GIC criteria, independent of the choice of the penalty multiplier.

Finally, we have restricted our discussion to particular (linear regression) models and a particular model selection method (GIC) for definiteness and simplicity but it is clear that the methodology can be extended to other cases. In particular, any model selection method which involves minimizing, over a set of models, a penalized function of the data in which the penalty is controlled by a penalty multiplier can be embedded in an appropriate model selection curve, and this opens up the possibility of using methods such as we have developed in this paper in other contexts.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium of Information Theory*, Eds. B.N. Petrov, F. Csáki, pp. 267–281. Akadémiai Kiadó: Budapest.

Bai, Z.D., Krishnaiah, P.R. & Zhao, L.C. (1986). On detection of the number of signals when the noise covariance matrix is arbitrary. *J. Multivariate Anal.*, **20**, 26–49.

Casella, G. & Consonni, G. (2009). Reconciling model selection and prediction. *arXiv:0903.3620v1*, 19 pp.

Claeskens, G. & Hjort, N.L. (2008). *Model selection and model averaging*. New York: Cambridge.

Flury, B. & Riedwyl, H. (1988). *Multivariate statistics. A practical approach*. London: Chapman and Hall.

Foster, D.P. & George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**, 1947–1975.

Gunst, G.P. & Mason, R.L. (1980). *Regression analysis and its applications*. New York: Marcel Dekker.

Hald, A. (1952). *Statistical theory with engineering applications*. New York: Wiley.

Hannan, E.J. & Quinn, B.G. (1979). The determination of the order of an autoregression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **41**, 190–195.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.

Hurvich, C.M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

Li, K.-C. (1987). Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.*, **15**, 958–975.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Miller, A. (2002). *Subset selection in regression*, 2nd ed. Boca Raton: Chapman & Hall/CRC.

Müller, S. & Welsh, A.H. (2005). Outlier robust model selection in linear regression. *J. Amer. Statist. Assoc.*, **100**, 1297–1310.

Müller, S. & Welsh, A.H. (2009). Robust model selection in generalized linear models. *Statist. Sinica*, **19**, 1155–1170.

Piepel, G. & Redgate, T. (1998). A mixture experiment analysis of the Hald cement data. *Amer. Statist.*, **52**, 23–30.

Rao, C.F. & Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.

Ronchetti, E., Field, C. & Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **92**, 1017–1023.

Ronchetti, E. & Staudte, R.G. (1994). A robust version of Mallows' C_P. *J. Amer. Statist. Assoc.*, **89**, 550–559.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Soc.*, **88**, 486–494.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, **91**, 655–665.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica*, **7**, 221–264.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43–49.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist. Theory Methods*, **A7**, 13–26.

Sugiyama, M. & Ogawa, H. (2001). Subspace information criterion for model selection. *Neural Comput.*, **13**, 1863–1889.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku*, **153**, 12–18.

Wisnowski, J.W., Simpson, J.R., Montgomery, D.C. & Runger, G.C. (2003). Resampling methods for variable selection in robust regression. *Comput. Statist. Data Anal.*, **43**, 341–355.

Woods, H., Steinour, H.H. & Starke, H.R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Indus. Eng. Chem.*, **24**, 1207–1214.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950.

Résumé

Beaucoup de méthodes populaires de sélection de variables impliquent la minimisation d'une fonction pénalisée des données (comme la vraisemblance maximisée ou la somme résiduelle carrés) sur un jeu de modèles. La pénalité dans la fonction de critère est contrôlée par un multiplicateur de pénalité λ qui détermine les propriétés de la procédure. Nous reconsidérons d'abord des critères de sélection modèles de la forme simple 'Perte + Pénalité' et proposons ensuite d'étudier de telles fonctions comme les fonctions du multiplicateur de pénalité. Cette approche peut être interprétée comme l'exploration de la stabilité de fonctions de critère par ce que nous appelons des courbes de choix modèles. Il mène à de nouvelles compréhensions dans le sélection de variables et de nouvelles propositions de la façon d'utiliser ces fonctions de critère pour sélectionner de variables. Nous utilisons le bootstrap pour augmenter des courbes de choix modèle et développent les résumés numériques et graphiques des résultats. La méthodologie est illustrée sur deux jeux de données et soutenue par une petite simulation. Nous montrons que la nouvelle méthodologie peut surpasser des méthodes comme AIC et BIC qui correspond aux points simples sur une courbe de choix modèle.

[*Received February 2009, accepted January 2010*]