Information criteria for variable selection under sparsity

MAARTEN JANSEN
Departments of Mathematics and Computer Science, Université Libre de Bruxelles
B-1050 Brussels, Belgium
maarten.jansen@ulb.ac.be

October 2013

Abstract

The optimization of an information criterion in a variable selection procedure leads to an additional bias, which can be substantial in sparse, high-dimensional data. The bias can be compensated by applying shrinkage while estimating within the selected models. This paper presents modified information criteria for use in variable selection and estimation without shrinkage. The analysis motivating the modified criteria follows two routes. The first, explored for signal-plus-noise observations only, goes by comparison of estimators with and without shrinkage. The second, discussed for general regression models, describes the optimization or selection bias as a double-sided effect, named a mirror effect in the paper: among the numerous insignificant variables, those with large, noisy values present themselves as being more valuable than an arbitrary variable, while in fact, they carry more noise than an arbitrary variable. The mirror effect is developed for Akaike's Information Criterion and for Mallows' C_p, with special attention to the latter criterion as a stopping rule in a least angle regression routine. The result is a new stopping rule, not focusing on the quality of a lasso shrinkage selection, but on the least squares estimator without shrinkage within the same selection.

Key words and phrases

High-dimensional data, Information Criterion, lasso, Mallows' C_p, Sparsity, Variable Selection.

1 Introduction

This paper presents information criteria for estimators without shrinkage in model selection. Although Mallows' C_p (Mallows, 1973) criterion is an unbiased estimator of the expected average squared prediction error of a model, it is an often reported fact (Woodroofe, 1982; Ishwaran, 2004; Loubes and Massart, 2004; Stine, 2004; Ye, 1998) that minimization of the criterion overestimates the number of variables needed to minimize the prediction error. Given an estimator within a selected model, Mallows' C_p, like many other information criteria, has the form of a penalized likelihood or sum of squared residuals. When the penalty depends on the model size, then among all models of equal size, selection is based on the sum of squared residuals. In the case of high-dimensional sparse models, it is easy to reduce the sum of squared residuals by a well-chosen combination of falsely significant variables, thereby fitting the observational errors. The false positives thus present themselves as being better in modelling the

observations than variables that were selected in a purely arbitrary way, whereas in reality their estimates deviate more from their values in the true model than do those for arbitrary variables. This two-sided effect of appearance versus reality can be described as a mirror effect, and is the topic of this paper.

The mirror effect can be seen as statistics of residuals that change through the optimization of an information criterion in variable selection. The outcome of the optimization depends on the errors, while an information criterion has been designed to evaluate the quality of one specific model. The change of statistics through the selection can be compensated for by a generalized concept of degrees of freedom (Ye, 1998), replacing the simple model size in the penalty. The mirror effect described in this paper is closely related to that concept.

The paper provides data-dependent expressions for penalties in information criteria that correct a priori for the mirror effect. In principle the mirror effect paradigm can be adopted with any distribution for the error, any set or search structure for the model selection problem, any information criterion and any estimator within the selected model. As the mathematical details depends on the case, most of the discussion in the paper concentrates on important examples, such as normal errors and least squares estimators. This paper discusses the application for both Mallows' C_p and Akaike's Information Criterion (Akaike, 1973). In the case of normal errors and Mallows' C_p, the resulting penalty term can be compared to a lower bound that avoids inconsistent estimators (Birgé and Massart, 2007). The mirror correction, being data-dependent, automatically finds the degree of sparsity in the given data. The simulation study in Section 2.6 illustrates that in terms of prediction error, the mirror correction slightly outperforms methods that control the false discovery rate (Benjamini and Hochberg, 1995) or even the absolute number of false positives (Donoho and Johnstone, 1994). These methods have been found to perform well in a minimax sense (Donoho and Johnstone, 1999) with respect to the prediction error, but the focus on false positives leads to estimators that are not adaptive to the true, significant components in the data.

The mirror correction proposed by this paper can also be seen as an alternative for shrinkage as a tool to compensate for optimization randomness. The idea behind shrinkage is to temper the effect of false positives. The tempering may even exactly undo the optimization bias. This occurs when the errors are normally distributed and the shrinkage is realized through ℓ_1 constrained regression, known as the lasso or least absolute shrinkage and selection operator (Tibshirani, 1996) or basis pursuit (Chen et al., 1998). Thanks to the shrinkage, the expression for Mallows' C_p in the optimization of the model uses the same penalty as for evaluation of an estimator without shrinkage in a fixed model. This penalty is based on the concept of generalized degrees of freedom (Ye, 1998). Both in low-dimensional (Zou et al., 2007) and in high-dimensional (Tibshirani and Taylor, 2012) data, the number of degrees of freedom during a lasso operation can be taken equal to model size. In the case of a signal-plus-noise model, the expression of Mallows' C_p thus reduces to that of Stein's unbiased risk estimator (Stein, 1981; Donoho and Johnstone, 1995; Loubes and Massart, 2004), while lasso itself becomes soft-thresholding.

Firstly, shrinkage thus reduces the effect of false positives. Secondly, it may also be superior to simple least squares in terms of prediction error, thanks to Stein's phenomenon (Stein, 1956). Thirdly, ℓ_1 regularized least squares is a convex optimization problem, as are variants such as the Dantzig selector (Candès and Tao, 2007). Without shrinkage, variable selection is a combinatorial optimization problem. Fourthly, for a given penalty value, ℓ_1 regularization imposes nearly the same degree of sparsity as an estimator penalized by the model size, without further shrinkage (Donoho, 2006). It has also been proved that, under certain conditions, ℓ_1 constrained optimization is variable selection consistent, provided that the true model variables are large enough, compared to the regularization parameter. That is, if all variables in the true model are sufficiently significant and if the regularization parameter is not too high,

then, for $n \to \infty$, the set of nonzero variables in the selection equals the true set with probability tending to one (Meinshausen and Bühlmann, 2006; Wainwright, 2009; Tropp, 2006; Zhao and Yu, 2006). Fifthly, as illustrated in Figure 1(b), when using shrinkage, the curvature of the prediction error as a function of model size is small near its minimum. This is in contrast to the delicate minimization of the prediction error in absence of shrinkage. Sixthly, shrinkage provides a continuous transition between selection and non-selection. Continuous operations are mathematically more tractable.

In spite of these benefits, the use of shrinkage may be problematic in high-dimensional problems. First, it introduces a bias in the estimated parameters, even if the parameter is highly significant. This can be controlled by choosing shrinkage rules that spare large variables (Gao, 1998), including Bayesian shrinkage (Johnstone and Silverman, 2004). Secondly, as shrinkage reduces the effect of false positives, it is tolerant to their presence. As a result, the shrinkage rule that minimizes the prediction error, rests on a model with too many nonzeros. The minimum with small curvature in Figure 1(b) confirms the illusion of an easy problem, whereas finding the best selection without shrinkage requires careful optimization. While ℓ_1 regularization mimics estimation without shrinkage quite well for fixed penalty values, the equivalence between ℓ_1 and estimation without shrinkage no longer holds for the optimization over the penalty, or, equivalently, the optimization over the model size. The rather poor behavior of shrinkage selection with data-driven choice of the penalty value explains why many state-of-the-art methods do not optimize over the regularization, but rather opt for a minimax choice of it.

2 Mirror effect in variable selection without shrinkage

2.1 Optimization bias

This paper investigates the selection of variables β_i in a regression model

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = K\boldsymbol{\beta} + \boldsymbol{\varepsilon} = K\boldsymbol{\beta} + \sigma\boldsymbol{Z}, \tag{1}$$

where $\boldsymbol{Z}$ is a n-dimensional vector of standardized, independent and identically distributed errors with $\mathrm{var}(Z_i) = 1$, for $i = 1, \ldots, n$. The design matrix K has thus n rows. The number of columns, m, may or may not be equal to n. In high-dimensional data, we typically find $m \gg n$, but we assume that the number of significant variables, n_1, is always smaller than n.

Let $\widehat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$ in model (1) where n_1 variables are allowed to be nonzero, and denote $\widehat{\boldsymbol{\mu}} = K\widehat{\boldsymbol{\beta}}$. The objective is to find the value of n_1 and the corresponding estimator $\widehat{\boldsymbol{\beta}}$ with n_1 nonzeros that minimizes the expected prediction error

$$\mathrm{PE}(\boldsymbol{x}) = \frac{1}{n} E\left(\|K\boldsymbol{\beta} - K\widehat{\boldsymbol{\beta}}\|^2\right). \tag{2}$$

The binary selection vector $\boldsymbol{x} \in \{0,1\}^m$ represents the model under consideration. Let $K_{\boldsymbol{x}}$ be the submatrix of K containing the columns corresponding to the 1's in $\boldsymbol{x}$. For any linear estimator $\widehat{\boldsymbol{\mu}} = A_{\boldsymbol{x}}\boldsymbol{Y}$ within a given, deterministic model $\boldsymbol{x}$, the prediction error is estimated unbiasedly by Mallows' C_p, which is in general $\Delta_p(A_{\boldsymbol{x}}) = n^{-1}\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}) + 2\sigma^2 n^{-1}\mathrm{tr}(A_{\boldsymbol{x}}) - \sigma^2$, where $\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}) = \|\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}\|^2$ is the sum of squared residuals. We use the symbol Δ_p because in most papers C_p stands for a standardized or studentized quantity. This paper concentrates on the least squares estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}} = (K_{\boldsymbol{x}}^T K_{\boldsymbol{x}})^{-1} K_{\boldsymbol{x}}^T \boldsymbol{Y}$, where $\boldsymbol{\beta}_{\boldsymbol{x}}$ denotes the subvector with the nonzero entries of $\boldsymbol{\beta}$ corresponding to the nonzeros in $\boldsymbol{x}$. Using

the orthogonality $K\widehat{\boldsymbol{\beta}} = K_{\boldsymbol{x}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}} \perp (\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}})$ it follows that $E\{\Delta_p(\boldsymbol{x})\} = n^{-1}(\|K\boldsymbol{\beta}\|_2^2 - E(\|K\widehat{\boldsymbol{\beta}}\|_2^2) + 2n_1\sigma^2)$. The only expectation in the expression for $E\{\Delta_p(\boldsymbol{x})\}$ can be rewritten as

$$E(\|K\widehat{\boldsymbol{\beta}}\|_2^2) = E(\|P_{\boldsymbol{x}}K\boldsymbol{\beta} + P_{\boldsymbol{x}}\boldsymbol{\varepsilon}\|_2^2), \tag{3}$$

where $P_{\boldsymbol{x}} = K_{\boldsymbol{x}}(K_{\boldsymbol{x}}^T K_{\boldsymbol{x}})^{-1}K_{\boldsymbol{x}}^T$ is the orthogonal projection onto the columns of $K_{\boldsymbol{x}}$.

In the case where the selection $\boldsymbol{X}$ depends on the observations, through optimization of $\Delta_p(\boldsymbol{x})$, (3) becomes

$$\begin{aligned} E(\|K\widehat{\boldsymbol{\beta}}\|_2^2) &= E(\|P_{\boldsymbol{X}}K\boldsymbol{\beta}\|_2^2) + E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2) + 2E\Big\{(P_{\boldsymbol{X}}K\boldsymbol{\beta})^T(P_{\boldsymbol{X}}\boldsymbol{\varepsilon})\Big\} \\ &= E(\|P_{\boldsymbol{X}}K\boldsymbol{\beta}\|_2^2) + E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2) + 2(K\boldsymbol{\beta})^T E(P_{\boldsymbol{X}}\boldsymbol{\varepsilon}). \end{aligned}$$

This leads to an expected value of Mallows' criterion, taking the optimization into account,

$$E_{\boldsymbol{X}}\{\Delta_p(\boldsymbol{X})\} = \frac{1}{n}\left\{\|K\boldsymbol{\beta}\|_2^2 - E(\|P_{\boldsymbol{X}}K\boldsymbol{\beta}\|_2^2) - E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2) - 2(K\boldsymbol{\beta})^T E(P_{\boldsymbol{X}}\boldsymbol{\varepsilon}) + 2n_1\sigma^2\right\}. \tag{4}$$

The expected prediction error, on the other hand, can be written as $\mathrm{PE}(\boldsymbol{x}) = n^{-1}\{\|K\boldsymbol{\beta}\|_2^2 + E(\|K\widehat{\boldsymbol{\beta}}\|_2^2) - 2(K\boldsymbol{\beta})^T E(K\widehat{\boldsymbol{\beta}})\}$. For fixed $\boldsymbol{x}$, the expressions of $\mathrm{PE}(\boldsymbol{x})$ and $E\{\Delta_p(\boldsymbol{x})\}$ lead to identical outcomes. For observation dependent selections $\boldsymbol{X}$,

$$\begin{aligned} E_{\boldsymbol{X}}\{\mathrm{PE}(\boldsymbol{X})\} &= \frac{1}{n}\left\{\|K\boldsymbol{\beta}\|_2^2 + E(\|P_{\boldsymbol{X}}K\boldsymbol{\beta}\|_2^2) + E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2) - 2(K\boldsymbol{\beta})^T E(P_{\boldsymbol{X}}K\boldsymbol{\beta})\right\} \\ &= \frac{1}{n}\left\{\|K\boldsymbol{\beta}\|_2^2 - E(\|P_{\boldsymbol{X}}K\boldsymbol{\beta}\|_2^2) + E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2)\right\}. \end{aligned} \tag{5}$$

The difference between (4) and (5) is due to the observation-dependent selection process, which is assumed to proceed in two steps. First, for a given model size n_1, the optimal n_1 term selection $\boldsymbol{X}_{n_1}$ is computed, where the optimization takes place over the observed values of the $\Delta_p(\boldsymbol{X})$ or of any other information criterion. Next, the prediction error of the best n_1 term approximation is considered as function of n_1.

The analysis of $E_{\boldsymbol{X}}\{\mathrm{PE}(\boldsymbol{X})\} - E_{\boldsymbol{X}}\{\Delta_p(\boldsymbol{X})\}$ is simplified by assuming that $2n^{-1}(K\boldsymbol{\beta})^T E(P_{\boldsymbol{X}}\boldsymbol{\varepsilon}) = o(n_1 n^{-1})$. In the signal-plus-noise case, for instance, this follows from a symmetric random model on $\boldsymbol{\beta}$, or from the sparsity in Assumption 2. This leads to an expression for the difference between (5) and (4), depending on n_1 only, $E_{\boldsymbol{X}}\{\mathrm{PE}(\boldsymbol{X}_{n_1})\} - E_{\boldsymbol{X}}\{\Delta_p(\boldsymbol{X}_{n_1})\} \approx 2m(n_1)$, where

$$m(n_1) = \frac{1}{n}E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2) - \frac{n_1}{n}\sigma^2. \tag{6}$$

As $\boldsymbol{X}$ can be observed, the selection bias can be estimated unbiasedly from

$$\widehat{m}(n_1) = \frac{1}{n}E\left(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2 \mid \boldsymbol{X}\right) - \frac{n_1}{n}\sigma^2. \tag{7}$$

The prediction error can then be estimated from $\widehat{\Delta}_p(\boldsymbol{X}_{n_1}) = \Delta_p(\boldsymbol{X}_{n_1}) + 2\widehat{m}(n_1)$.

This paper further analyzes the bias correction $2m(n_1)$, describing it in terms of an oracular variable selection, defined as follows.

Definition 1 *Given the model (1), define the class of submodels $\mathcal{M} \subseteq \{0,1\}^m$, using the binary representation introduced above. For each submodel $\boldsymbol{x} \in \mathcal{M}$, consider the least squares estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}}$ within that model.*

Then the oracle selection is the model $\boldsymbol{x}^o_{n_1}$ that minimizes $\lim_{\sigma\to 0} \mathrm{PE}(\boldsymbol{x})$ among all models in $\mathcal{M}$ with size n_1. In other words, it is the output from a model selection and estimation that has $K\boldsymbol{\beta}$ as input, rather than $\boldsymbol{Y} = K\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Given n_1 and the oracle selection $\boldsymbol{x}^o_{n_1}$, its least squares prediction error $\mathrm{PE}(\boldsymbol{x}^o_{n_1})$ is the mirror function.

The mirror function is thus the prediction error of a routine whose model selection is based on the oracular observations $K\boldsymbol{\beta}$, whereas its estimator within the selected model is based on the observations $\boldsymbol{Y} = K\boldsymbol{\beta}+\boldsymbol{\varepsilon}$. As the selection $\boldsymbol{x}^o_{n_1}$ does not depend on $\boldsymbol{\varepsilon}$, $E\{\Delta_p(\boldsymbol{x}^o_{n_1})\} = \mathrm{PE}(\boldsymbol{x}^o_{n_1})$. The use of the term mirror function is motivated by the following argument. Again under the mild assumptions of sparsity, stated in Section 2.3, it holds that $E_{\boldsymbol{X}}\{\mathrm{PE}(\boldsymbol{X}_{n_1})\} - \mathrm{PE}(\boldsymbol{x}^o_{n_1}) \approx m(n_1) \approx \mathrm{PE}(\boldsymbol{x}^o_{n_1}) - E_{\boldsymbol{X}}\{\Delta_p(\boldsymbol{X}_{n_1})\}$. The oracle prediction error thus acts as the mirror that reflects $\Delta_p(\boldsymbol{X}_{n_1})$ onto $\mathrm{PE}(\boldsymbol{X}_{n_1})$ and vice versa.

2.2 The mirror and other penalties

Defining the residual vector $\boldsymbol{e} = \boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}$ and the generalized degrees of freedom (Ye, 1998) $\nu(n_1) = E\{\boldsymbol{\varepsilon}^T(\boldsymbol{\varepsilon} - \boldsymbol{e})\}\sigma^{-2}$, it is well known that $\Lambda_p(\boldsymbol{X}_{n_1}) = \mathrm{SS_E}(\widehat{\boldsymbol{\beta}}) + 2\nu(n_1)\sigma^2 n^{-1} - \sigma^2$ is an unbiased estimator of $E_{\boldsymbol{X}}\{\mathrm{PE}(\boldsymbol{X}_{n_1})\}$, for any choice of $\boldsymbol{X}_{n_1}$, random or fixed. The approximation proposed in Section 2.1, (6), can thus be written as $\nu(n_1) = E(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2)\sigma^{-2} + o(n_1)$ and, consequently, $m(n_1) = \{\nu(n_1) - n_1\}n^{-1}\sigma^2 + o(n_1 n^{-1})$.

The mirror corrected penalty can be compared to the minimum penalty for consistent estimators (Birgé and Massart, 2007). Being a lower bound, that penalty is not data-specific, unlike that proposed in this paper. The same remark holds for the penalties proposed in Abramovich et al. (2007), for instance. Simulations discussed in the Supplementary Material show that the mirror penalty detects the degree of sparsity automatically. It can be shown that for n_1 larger than that degree, the mirror penalty $\nu(n_1)$ increases faster than the lower bound of Birgé and Massart (2007). Unlike that lower bound, however, the mirror paradigm is not limited to normal errors or to Mallows' C_p criterion. See the Supplementary Material for a full discussion.

2.3 Signal-plus-noise, using a random model for $\boldsymbol{\beta}$

We start the study of (6) in a simple signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the sparse signal $\boldsymbol{\beta}$ is observed directly, and $m = n$. Extension to the general form of (1) follows in Section 4. The least squares estimator for given $\boldsymbol{x}$ is $\widehat{\beta}_i = Y_i\, x_i$, where x_i is a component of the selection vector $\boldsymbol{x}$. The best n_1 term selection, measured by the C_p-value, consists of the n_1 largest elements from $\boldsymbol{Y}$.

The study is facilitated by assuming that the sparse vector of parameters $\boldsymbol{\beta}$ constitute an n-tuple of independent realizations from a random variable β_n with a density function $f_{\beta_n}(v)$. The subscript n denotes dependence on n, which will allow us to impose increasing sparsity in an asymptotic analysis. The eventual outcome will be independent of the precise form of $f_{\beta_n}(v)$.

In the signal-plus-noise model $\boldsymbol{Y}_n = \boldsymbol{\beta}_n + \boldsymbol{\varepsilon}$, the error distribution is assumed to be independent from n with variance $\sigma^2 = E(\varepsilon^2)$.

We let $\mathcal{X}_{n_1}$ denote the active subset of the index set $\{1,\ldots,n\}$, corresponding to the ones in the binary vector $\boldsymbol{X}_{n_1}$. The functions $\Delta_p(\mathcal{X}_{n_1})$ and $\mathrm{PE}(\mathcal{X}_{n_1})$ will be used to denote $\Delta_p(\boldsymbol{X}_{n_1})$ and $\mathrm{PE}(\boldsymbol{X}_{n_1})$. We let $\mathcal{X}'_{n_1}$ denote the complement of $\mathcal{X}_{n_1}$ in $\{1,\ldots,n\}$. The set $\mathcal{X}'_{n_1}$ contains the indices of the variables with the $n_0 = n - n_1$ smallest magnitudes. Defining the event $S_{n,n_0} = \{$ In a set of n independent, identically distributed realizations, the observed $|Y_n|$ is among the n_0 smallest magnitudes$\}$, we have $P(S_{n,n_0}) = n_0 n^{-1}$. Symmetry in the random model for β_n then allows us to state that $E\{\Delta_p(\mathcal{X}_{n_1})\} = n_0 n^{-1} E(Y_n^2 \mid S_{n,n_0}) + 2n_1 n^{-1}\sigma^2 - \sigma^2$. We also define the oracular version of the event S_{n,n_0} as $O_{n,n_0} = \{$ In a set of n independent, identically distributed realizations, the observed $|\beta_n|$ is among the n_0 smallest magnitudes$\}$. The complement of O_{n,n_0} corresponds to the selection $\boldsymbol{x}^o_{n_1}$ in Definition 1. Let $\mathcal{X}^o_{n_1}$ be the set of indices i for which $x^o_{n_1,i} = 1$. Starting from $E(Y_n^2 \mid O_{n,n_0}) = \sigma^2 + E(\beta_n^2 \mid O_{n,n_0})$ it follows that $E\{\Delta_p(\mathcal{X}^o_{n_1})\} = n_0 n^{-1} E(Y_n^2 \mid O_{n,n_0}) + 2n_1 n^{-1}\sigma^2 - \sigma^2 = n_1 n^{-1}\sigma^2 + n_0 n^{-1} E(\beta_n^2 \mid O_{n,n_0})$ and thus $E\{\Delta_p(\mathcal{X}^o_{n_1})\} - E\{\Delta_p(\mathcal{X}_{n_1})\} = n_0 n^{-1}\{\sigma^2 + E(\beta_n^2 \mid O_{n,n_0}) - E(Y_n^2 \mid S_{n,n_0})\}$. A mirrored relation holds between the prediction errors. In order to check this, we start from a conditioning of the prediction error on O_{n,n_0} to find that $\mathrm{PE}(\mathcal{X}^o_{n_1}) = n_1 n^{-1}\sigma^2 + n_0 n^{-1} E(\beta_n^2 \mid O_{n,n_0})$, in line with the unbiasedness of $\Delta_p(\mathcal{X}^o_{n_1})$. The prediction error can be written as

$$\begin{aligned}
\mathrm{PE}(\mathcal{X}_{n_1}) &= \mathrm{PE}(\mathcal{X}_{n_1} \mid S_{n,n_0})\, P(S_{n,n_0}) + \mathrm{PE}(\mathcal{X}_{n_1} \mid S'_{n,n_0})\, P(S'_{n,n_0}) \\
&= E(\beta_n^2 \mid S_{n,n_0})\,\frac{n_0}{n} + E(\varepsilon^2 \mid S'_{n,n_0})\,\frac{n_1}{n} = \frac{n_0}{n} E(\beta_n^2 \mid S_{n,n_0}) + \sigma^2 - \frac{n_0}{n} E(\varepsilon^2 \mid S_{n,n_0}) \\
&= \frac{n_1}{n}\sigma^2 + \frac{n_0}{n}\Big\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0}) + E(\beta_n^2 \mid S_{n,n_0})\Big\}. \qquad (8)
\end{aligned}$$

We now impose that the vector of β_n is sparse enough to allow an asymptotically perfect separation between significant and error-dominated variables:

Assumption 1 *When $n \to \infty$, the prediction error of an oracular component selection is dominated by the error present in the observations of the selected variables, that is* $\mathrm{PE}(\mathcal{X}^o_{n_1}) = E\{\Delta_p(\mathcal{X}^o_{n_1})\} \sim n_1 n^{-1}\sigma^2$.

An implication of Assumption 1 follows from the above stated expression of $\mathrm{PE}(\mathcal{X}^o_{n_1})$. We find $n_1 n^{-1}\sigma^2 + n_0 n^{-1} E(\beta_n^2 \mid O_{n,n_0}) \sim n_1 n^{-1}\sigma^2$, which becomes $E(\beta_n^2 \mid O_{n,n_0}) = o(n_1 n^{-1})$.

The following assumption is about the performance of the non-oracular selection method.

Assumption 2 *The selection S_{n,n_0} performs asymptotically as well as O_{n,n_0}, in the sense that $E(\beta_n^2 \mid S_{n,n_0}) = o(n_1 n^{-1})$. as $n \to \infty$.*

In terms of a non-random model for $\boldsymbol{\beta}_n$, this means that the threshold λ_{n_1} selecting n_1 significant variables satisfies $n^{-1}\sum_{i=1}^n \beta_i^2 P(|Y_i| < \lambda) = o(n_1 n^{-1})$. The Supplementary Material includes a quantitative discussion of the interpretation of Assumption 2 in function spaces imposing sparsity, such as ℓ_p balls with $p < 2$ or multiscale sparsity, such as Besov spaces. The discussion involves the introduction of an index of sparsity, inspired by the g-index from bibliometry (Egghe, 2006). Assumption 2 is satisfied if the data vector $\boldsymbol{\beta}_n$ is sparse, if the noise is not heavy tailed, so that it can be easily separated from the data, and if the threshold or model size is near its optimal value.

Assumption 2 implies that

$$0 \le E(Y_n^2 \mid S_{n,n_0}) - E(\varepsilon^2 \mid S_{n,n_0}) = o\left(\frac{n_1}{n}\right). \qquad (9)$$

This follows from the equation $E(Y_n^2 \mid S_{n,n_0}) - E(\varepsilon^2 \mid S_{n,n_0}) = E(\beta_n^2 \mid S_{n,n_0}) + 2E(\varepsilon\beta_n \mid S_{n,n_0})$, and the fact that $E(\varepsilon\beta_n \mid S_{n,n_0}) < 0$.

Assumptions 1 and 2 allow us to conclude that approximating $\mathrm{PE}(\mathcal{X}_{n_1})$ as the reflection of $E\{\Delta_p(\mathcal{X}_{n_1}\}$ with respect to the oracular mirror $E\{\Delta_p(\mathcal{X}^o_{n_1})\} = \mathrm{PE}(\mathcal{X}^o_{n_1})$ does not disturb the optimization of the prediction error. More precisely, introduce the approximation errors $\Delta_{1,n}$ and $\Delta_{2,n}$ by

$$\begin{aligned} E\{\Delta_p(\mathcal{X}^o_{n_1})\} - E\{\Delta_p(\mathcal{X}_{n_1})\} &= \frac{n_0}{n}\Big\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0})\Big\} + \Delta_{1,n}, \\ \mathrm{PE}(\mathcal{X}_{n_1}) - \mathrm{PE}(\mathcal{X}^o_{n_1}) &= \frac{n_0}{n}\Big\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0})\Big\} + \Delta_{2,n}. \end{aligned}$$

Then $\lim_{n\to\infty} q_n(n_1) = 0$, where $q_n(n_1) = \Delta_n/\mathrm{PE}(\mathcal{X}_{n_1})$ and $\Delta_n = \Delta_{1,n}+\Delta_{2,n}$. Defining $\mathrm{PE}_\Delta(\mathcal{X}_{n_1}) = \mathrm{PE}(\mathcal{X}_{n_1}) - \Delta_n$, we have, for $n \to \infty$ and any n_1, that $-q_n(n_1)\mathrm{PE}(\mathcal{X}_{n_1}) \le \mathrm{PE}_\Delta(\mathcal{X}_{n_1}) - \mathrm{PE}(\mathcal{X}_{n_1}) \le q_n(n_1)\mathrm{PE}(\mathcal{X}_{n_1})$, or, equivalently, $\mathrm{PE}(\mathcal{X}_{n_1})\{1 - q_n(n_1)\} \le \mathrm{PE}_\Delta(\mathcal{X}_{n_1}) \le \mathrm{PE}(\mathcal{X}_{n_1})\{1 + q_n(n_1)\}$. So, if $\widehat{n}_1$ and $\widetilde{n}_1$ optimize $\mathrm{PE}(\mathcal{X}_{n_1})$ and $\mathrm{PE}_\Delta(\mathcal{X}_{n_1})$ respectively, then

$$\Big\{1 - q_n(\widetilde{n}_1)\Big\}\mathrm{PE}(\mathcal{X}_{\widetilde{n}_1}) \le \mathrm{PE}_\Delta(\mathcal{X}_{\widetilde{n}_1}) \le \mathrm{PE}_\Delta(\mathcal{X}_{\widehat{n}_1}) \le \Big\{1 + q_n(\widehat{n}_1)\Big\}\mathrm{PE}(\mathcal{X}_{\widehat{n}_1})$$

or

$$1 \le \frac{\mathrm{PE}(\mathcal{X}_{\widetilde{n}_1})}{\mathrm{PE}(\mathcal{X}_{\widehat{n}_1})} \le \frac{1 + q_n(\widehat{n}_1)}{1 - q_n(\widetilde{n}_1)}. \tag{10}$$

Thus the minimizers of the exact and approximate prediction errors have asymptotically the same efficiency with respect to the prediction error. The approximate prediction error in its turn is estimated unbiasedly by $\widetilde{\Delta}_p(\mathcal{X}_{n_1}) = \Delta_p(\mathcal{X}_{n_1}) + 2m(n_1)$, with

$$\begin{aligned} m(n_1) &= \frac{n_0}{n}\Big\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0})\Big\} = P(S_{n,n_0})\Big\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0})\Big\} \qquad (11) \\ &= \int_{-\infty}^{\infty} f_{\beta_n}(v) \int_{-\infty}^{\infty} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y_n = v + e)\, de\; dv. \end{aligned}$$

The mirror (11) and the corresponding double correction are illustrated in Figure 1(a), which depicts the apparent information for a given model size n_1, found by minimizing Mallows' C_p, along with the minimum prediction error for that model size. The contradiction between better-than-average appearance and worse-than-average reality is seen in the two curves being reflections of each other with respect to the oracular curve. The C_p curve has a minimum with small curvature, creating the illusion of an easy problem. The model selected using this curve is however far too large.

2.4 The mirror effect in terms of thresholds

In this section we seek approximations to the mirror effect that satisfy three conditions. Firstly, the error of approximation is small compared to the prediction error, in the sense that, asymptotically, it does not disturb optimization of the estimated prediction error curve. Secondly, the expression is easy to implement. Thirdly, for normal errors, it reduces to an expression that can be derived as a hard threshold correction of Stein's unbiased risk estimator. This correction is further discussed in Section 3.

We define the expected mirror contribution for a given component value v as

$$t(n_1, v) = \int_{-\infty}^{\infty} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y = v + e)\, de. \tag{12}$$

The expected mirror in the model is then

$$m(n_1) = \int_{-\infty}^{\infty} f_{\beta_n}(v) t(n_1, v)\, dv.$$

In a similar way, we can write the contribution of one component to the expected prediction error, given its value v, as

$$r(n_1, v) = v^2\, P(S_{n,n_0} \mid \beta_n = v) + E(\varepsilon^2 \mid S'_{n,n_0})\, P(S'_{n,n_0} \mid \beta_n = v). \tag{13}$$

The following lemma, proved in Appendix A, states that the expected mirror can be approximated by assuming for each individual component that its error-free value is zero. The approximation error, relative to the prediction error, tends to zero.

Lemma 1 *Suppose that we observe n independent samples from $Y_n = \beta_n + \varepsilon$, with β_n and ε independent. Further assume that the distributions of ε and β_n are symmetric around the origin, and that ε has a unimodal distribution and a quantile function satisfying $Q_\varepsilon(1) = \infty$. We impose the following conditions:*

1. *the density $f_\varepsilon(e)$ has a bounded second derivative;*
2. *the density $f_\varepsilon(e)$ shows exponential decay as $|e| \to \infty$;*
3. *the large values of β_n dominate the errors. More precisely, the decay of $f_\varepsilon(e)$ is essentially faster than that of $f_{Y_n}(e)$ in the sense that*

$$\lim_{e \to \pm\infty} \frac{\log f_\varepsilon(e)}{\log f_{Y_n}(e)} = \infty;$$

4. *The large values of β_n are sparse, in the sense that there exists a positive p^* so that for any positive δ one can find an integer n^* for which $P(|\beta_n| < \delta) \geq p^*$, for any integer $n \geq n^*$.*

Further assume that $n_1/n \to 0$ as $n \to \infty$. Then the function $t(n_1, v)$ defined in (12) satisfies

$$\lim_{n \to \infty} \frac{t(n_1, \beta_n) - t(n_1, 0)}{r(n_1, \beta_n)} = 0, \tag{14}$$

for any sequence β_n. Hence

$$\lim_{n \to \infty} \frac{m(n_1) - t(n_1, 0)}{r(n_1, \beta_n)} = 0.$$

We can thus use $t(n_1, 0)$ use as an approximate mirror.

In a final step we further approximate the mirror by replacing the $P(S_{n,n_0} \mid Y_n = e)$ by a binary function $I(|u| < \lambda_{n_1})$, with an appropriate threshold λ_{n_1}.

Lemma 2 *Defining the threshold $\lambda_{n_1} = Q_{|Y_n|}(n_0 n^{-1})$, where $Q_{|Y_n|}$ is the quantile function of $|Y_n|$, and*

$$\tau(\lambda_{n_1}) = \int_{-\lambda_{n_1}}^{\lambda_{n_1}} (\sigma^2 - e^2) f_\varepsilon(e) de, \tag{15}$$

then, if $n_0/n \to 1$ for $n \to \infty$, and if the error-free data are sparse and dominant in the sense of Lemma 1, $\lim_{n\to\infty} \{m(n_1) - \tau(\lambda_{n_1})\} / r(n_1, \beta_n) = 0$.

The proof is in Appendix A.

An argument similar to that in (10) ensures that replacing $n_0 n^{-1}\left\{\sigma^2 - E(\varepsilon^2 \mid S_{n,n_0})\right\}$ with its approximation does not disturb the minimization of $E\{\Delta_p(\mathcal{X}_{n_1})\}$. Referring to the discussion of (11), the mirror correction can thus safely be approximated as $\mathrm{PE}(\mathcal{X}_{n_1}) - E\{\Delta_p(\mathcal{X}_{n_1})\} \sim 2\tau(\lambda_{n_1})$.

This expression does not depend on a model for β_n, except through the threshold λ_{n_1}. This threshold can, however, be easily replaced by the empirical value $\widehat{\lambda}_{n_1} = |Y|_{(n-n_1:n)}$, where $|Y|_{(n-n_1:n)}$ stands for the $(n-n_1)$th order statistic in an n-vector $\boldsymbol{Y}$.

If $\varepsilon \sim N(0, \sigma^2)$, then the correction reduces to

$$\mathrm{PE}(\mathcal{X}_{n_1}) - E\{\Delta_p(\mathcal{X}_{n_1})\} \sim 4\sigma^2\, \lambda_{n_1} \phi_\sigma(\lambda_{n_1}), n \to \infty, \tag{16}$$

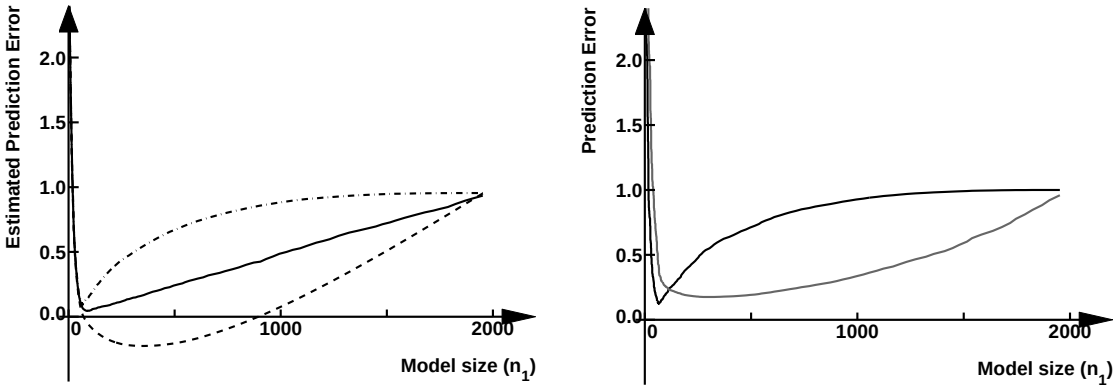
where $\phi_\sigma(e)$ is the density of zero mean normal random variable with variance σ^2.

2.5 Illustration of the mirror effect

The simulation in Figure 1 illustrates the discussions of the preceding sections. It was set up as follows. A vector of $n = 2000$ sparse data $\boldsymbol{\beta}$ was generated according to the zero inflated Laplace model $f_{\beta|\beta\neq 0}(v) = (a/2)\exp(-a|\beta|)$, where, in this simulation, $a = 1/5$ and $P(\beta \neq 0) = 1/20$. The observations are $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of independent, standard normal errors. For this model, the figure depicts the curve of $\Delta_p(\boldsymbol{X}_{n_1})$ as a function of n_1. As defined in Section 2.1, $\boldsymbol{X}_{n_1}$ is the n_1 term selection that minimizes $\Delta_p(\boldsymbol{X})$. For the same selection, Figure 1(a) also plots $\mathrm{PE}(\boldsymbol{X}_{n_1})$. The same plot contains the mirror curve $\mathrm{PE}(\boldsymbol{x}^o_{n_1})$, defined in Definition 1. Finally, Figure 1(b) contains the curve of $\Delta_p(\boldsymbol{X}_{n_1})$ when using soft-threshold shrinkage within the models $\boldsymbol{X}_{n_1}$.

2.6 A comparative simulation study in the signal-plus-noise model

The simulation study, summarized in Table 1, compares the efficiency of several methods for sparse variable selection with respect to the oracular prediction error, that is, $\mathrm{Eff} = \mathrm{PE}(\mathrm{oracle})/\mathrm{PE}$. The oracle would select all variables with error-free value above the noise standard deviation σ. The data were generated as in Johnstone and Silverman (2004), except for the sample size, which was taken to be $n = 10,000$ instead of $n = 1000$. One hundred replications of a n-vector of observations $\boldsymbol{Y}$ were generated, where $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The error vector $\boldsymbol{\varepsilon}$ is independent, homoscedastic, and normally distributed, whereas the error-free data $\boldsymbol{\beta}$ are set to zero, except for a proportion p of the variables, whose values are μ_0. The sparsity parameter p equals $p = 0.005$, while $\mu_0 = 7$. The table confirms the relatively low efficiencies, reported in Johnstone and Silverman (2004), of soft threshold methods using thresholds that estimate the minimum prediction error. The poor performance is entirely due to the oversmoothing of soft-thresholding. Indeed, hard thresholding focussing on the false discovery rate (Benjamini and Hochberg, 1995) or using empirical Bayes posterior median thresholds (Johnstone and Silverman, 2004) is outperformed by hard thresholding minimizing generalized cross validation, which estimates the prediction error. Indeed, its observed median efficiency is higher, as is its 95% quantile. The lower 5% efficiencies are, however, slightly less favorable for generalized cross validation than for the false discovery rate and empirical Bayes methods. Closer inspection of the simulation study, not shown in this table, reveals that this is due to imperfect estimation of the prediction error using generalized cross validation. These imperfections are a drawback for any method that estimates the the prediction error in a direct, data-adaptive way, rather than relying on minimax results (Donoho and Johnstone, 1994, 1999).



(a) (b)

Figure 1: Mallows's C_p in sparse variable selection with and without shrinkage. (a) Mirror effect, defined in (11). The dashed line depicts Mallows's C_p for the selections $\mathcal{X}_{n_1}$ that minimize $\Delta_p(\mathcal{X})$, given model sizes n_1. The dash-dot line represents the prediction errors $\mathrm{PE}(\mathcal{X}_{n_1})$ for the same selections. The curves of Mallows's C_p and the prediction errors are reflections of each other with respect to the oracular curve $\mathrm{PE}(\mathcal{X}^o_{n_1})$, depicted as a solid line. That mirror curve is the prediction error for a selection based on the error-free values. (b) Prediction errors for hard- and soft-thresholding, black and grey lines respectively. The hard threshold curve is the same as the dash-dot line in (a).

The table also illustrates that generalized cross validation is a more robust estimator of the prediction error than is Stein's unbiased risk estimator.

3 Undoing soft threshold bias

3.1 Soft-thresholding and Stein's unbiased risk estimator

This section shows that, for the case of normal errors, the correction term for Mallows' C_p in (16) can be obtained from an analysis that imports the difference between soft- and hard-threshold prediction errors into the expression of Stein's unbiased risk estimator.

Given a threshold value λ, the difference in prediction errors between soft- and hard-thresholding equals

$$\begin{aligned}\mathrm{PE}(\widehat{\boldsymbol{\beta}}^{\mathrm{H}}_{\lambda}) - \mathrm{PE}(\widehat{\boldsymbol{\beta}}^{\mathrm{S}}_{\lambda}) &= -\frac{\lambda^2}{n}\sum_{i=1}^{n} P(|Y_i| > \lambda) + \frac{2\lambda}{n}\sum_{i=1}^{n} E(\varepsilon_i X_i^+) - \frac{2\lambda}{n}\sum_{i=1}^{n} E(\varepsilon_i X_i^-) \\ &= -\frac{\lambda^2}{n}\sum_{i=1}^{n} P(X_i = 1) + \frac{2\lambda}{n}\sum_{i=1}^{n} E\left\{\mathrm{sign}(\beta_i + \varepsilon)\,\varepsilon X_i\right\}, \qquad (17)\end{aligned}$$

where $X^+ = I(Y > \lambda)$ and $X^- = I(Y < -\lambda)$. The first term of (17) can be estimated unbiasedly by $-\lambda^2 n^{-1}\sum_{i=1}^{n} X_i = -\lambda^2 N_1 n^{-1}$, where N_1 is the total number of observed magnitudes above the threshold.

	5%	50%	95%
SURE-soft	9.1	12.8	17.8
GCV-soft	8.5	12.9	17.9
EBayesthresh	36.4	58.9	88.9
FDR-thresh	35.0	58.8	100.0
SURE-hard	27.4	48.1	92.6
GCV-hard	34.4	73.4	100.0

Table 1: Quantiles of observed efficiencies in percentages for several threshold methods: SURE stands for Stein's unbiased risk estimation, GCV for generalized cross validation, FDR for false discovery rate; soft and hard stand for soft- and hard-thresholding.

The second term (17) cannot be estimated in an unbiased way. It can, however, be approximated by a constant, dependent on the threshold value, but not on $\boldsymbol{\beta}$. As follows from Proposition 1, the approximation error tends to zero more rapidly than the prediction error itself, so it does not disturb the maximization of the prediction error or of any estimate of it.

Proposition 1 *Let ε be symmetrically distributed with an exponentially decaying density function $f_\varepsilon(e)$ for which $\lim_{e\to\pm\infty} f'_\varepsilon(e)$ exists, and let*

$$\kappa(\lambda,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\mathrm{sign}(\beta_i+\varepsilon)I(|\beta_i+\varepsilon|>\lambda)\varepsilon\right\}. \tag{18}$$

Then there exists a function $c(\lambda)$ such that for any parameter vector $\boldsymbol{\beta}$

$$\lambda\,\frac{|\kappa(\lambda,\boldsymbol{\beta})-\kappa(\lambda,0)|}{\mathrm{PE}(\widehat{\boldsymbol{\beta}}^H_\lambda)} \le c(\lambda),\ \textit{with}\ \lim_{\lambda\to\infty} c(\lambda) = 0. \tag{19}$$

The proof is established in the Supplementary Material; see also Section 4.3.

An argument similar to that in (10) allows us to replace $\kappa(\lambda,\boldsymbol{\beta})$ by $\kappa(\lambda,\mathbf{0}) = E\left\{|\varepsilon|\,I(|\varepsilon|>\lambda)\right\}$ while keeping the quality of the minimization of $\mathrm{PE}(\widehat{\boldsymbol{\beta}}^H_\lambda)$.

In the case of soft thresholding, the well known (Stein, 1981; Donoho and Johnstone, 1995) expression for unbiased risk estimation for data with normally distributed errors is $\mathrm{SURE}(\widehat{\boldsymbol{\beta}}_\lambda) = n^{-1}\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_\lambda) + 2N_1 n^{-1}\sigma^2 - \sigma^2$. A quasi unbiased estimator for the hard thresholding prediction error can be obtained by adding the estimator $-\lambda^2 N_1 n^{-1} + 2\kappa(\lambda,0)$ for the terms of (17) to Stein's unbiased risk estimator. It is straightforward to verify that $\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{ST}}}) = \mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{HT}}}) + \lambda^2 N_1 n^{-1}$. Moreover, for normal errors $\varepsilon \sim N(0,\sigma^2)$, we have that $E(\varepsilon X^+;0) = \sigma^2\lambda\phi_\sigma(\lambda) = \sigma^2(\lambda/\sigma)\phi_1(\lambda/\sigma)$, leading to

$$\mathrm{SURE_H}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{HT}}}) = \mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{HT}}}) + \frac{2N_1}{n}\sigma^2 - \sigma^2 + 4\sigma^2\lambda\phi_\sigma(\lambda). \tag{20}$$

This is the same expression as (16), which followed from a different strategy and different approximations. The strategy in Sections 2.3 and 2.4 was first to quantify the mirror effect and then to approximate

it using a threshold expression, leading to (16). The current section has started from the observation that soft-threshold shrinkage perfectly compensates for the mirror effect in the case of normal errors. From there, approximating the difference between soft- and hard-thresholding has led to (20).

The expression can be further expanded towards generalized cross validation for hard thresholding.

3.2 Akaike's information criterion

For a given selection $\boldsymbol{x}$, Akaike's information criterion can be defined as $\mathrm{AIC}(\boldsymbol{x}) = 2\log\mathcal{L}(\boldsymbol{x}) - 2n_1 n^{-1}$, where $\mathcal{L}(\boldsymbol{x})$ is the maximum likelihood value within selection $\boldsymbol{x}$. This criterion is an asymptotically unbiased estimator for the Kullback–Leibler distance between a given model and the true observational distribution. When the criterion is used for optimization, the mirror correction in the case of signal-plus-noise observations can be found with similar arguments as for Mallows' criterion to be

$$\mathrm{AIC}^m(n_1) = -\frac{n_0}{n} - \log(2\pi\widehat{\sigma}^2) - 2\frac{n_1}{n}\frac{\sigma^2}{\sigma^2_\lambda} - 2\frac{\widetilde{n}_0}{n}\left(\frac{\sigma^2}{\sigma^2_\lambda} - 1\right), \tag{21}$$

where, for $\lambda = \lambda_{n_1}$ as defined in Lemma 2, $\sigma^2_\lambda = E(\varepsilon^2 | -\lambda_{n_1} < Y_n < \lambda_{n_1})$ and $\widetilde{n}_0 = n\,P(|\varepsilon| < \lambda_{n_1})$. The variance estimator $\widehat{\sigma}^2 = n_0^{-1}\sum_{i=1}^n (Y_i - \widehat{\beta}_i)^2$ is based on squared residuals within the model under consideration, while the variance σ^2 itself, in practice, is estimated in a way independent from the model, or at least in a robust way, such as using the median absolute deviation. An alternative variance estimator, based on generalized cross validation, is reported to be more robust, leading to better estimates of the Kullback–Leibler distance.

For normal observations, the criterion reduces to

$$\mathrm{AIC}^m(n_1) = -\frac{n_0}{n} - \log(2\pi\widehat{\sigma}^2) - 2\frac{\widetilde{n}_0}{n}\left\{\frac{n_1/n + 2\lambda_{n_1}\phi_\sigma(\lambda_{n_1})}{\widetilde{n}_0/n - 2\lambda_{n_1}\phi_\sigma(\lambda_{n_1})}\right\}. \tag{22}$$

The mirror effect in Akaike's criterion is discussed in the Supplementary Material.

4 The mirror effect in sparse regression models

4.1 The mirror effect on covariance matrices

For the development of (7) for the mirror estimator in the general regression model (1), we define $\boldsymbol{\eta} = K^T\boldsymbol{\varepsilon}$ with covariance matrix $\Sigma_{\boldsymbol{\eta}} = K^T K\sigma^2$. Then $E\left(\|P_{\boldsymbol{X}}\boldsymbol{\varepsilon}\|_2^2 \mid \boldsymbol{X}\right) = \sigma^2 E(\boldsymbol{\eta}_{\boldsymbol{X}}^T\Sigma^{-1}_{\boldsymbol{\eta},\boldsymbol{X}\boldsymbol{X}}\boldsymbol{\eta}_{\boldsymbol{X}} \mid \boldsymbol{X})$, where $\Sigma_{\boldsymbol{\eta},\boldsymbol{X}\boldsymbol{X}}$ is the submatrix of $\Sigma_{\boldsymbol{\eta}}$ with the rows and columns corresponding to the 1's in $\boldsymbol{X}$. In a similar way, submatrices are defined for the 1's in the complementary binary vector $\boldsymbol{X}' = \boldsymbol{1} - \boldsymbol{X}$.

Writing

$$E(\boldsymbol{\eta}_{\boldsymbol{X}}^T\Sigma^{-1}_{\boldsymbol{\eta},\boldsymbol{X}\boldsymbol{X}}\boldsymbol{\eta}_{\boldsymbol{X}} \mid \boldsymbol{X}=\boldsymbol{x}) = \mathrm{tr}(\Sigma^{-1}_{\boldsymbol{\eta},\boldsymbol{x}\boldsymbol{x}}\Sigma_{\boldsymbol{\eta}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{x}\boldsymbol{x}}) + E(\boldsymbol{\eta}_{\boldsymbol{x}}^T \mid \boldsymbol{X}=\boldsymbol{x})\Sigma^{-1}_{\boldsymbol{\eta},\boldsymbol{x}\boldsymbol{x}}E(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{X}=\boldsymbol{x}),$$

the second term is zero if we again consider a symmetric, random model for $\boldsymbol{\beta}$, so that the selection event $\{\boldsymbol{X} = \boldsymbol{x}\}$ preserves the symmetry in the error distribution. The remainder of this section concentrates on the first term, which is the trace of a product of two matrices. The first, $\Sigma^{-1}_{\boldsymbol{\eta},\boldsymbol{x}\boldsymbol{x}}$, is an inverse submatrix of the unconditional covariance matrix. The second matrix has the same rows and columns, indicated by $\boldsymbol{x}$, but this time taken from the matrix of conditional covariances for the selection event.

The distribution of a fully unconditional quadratic form $\boldsymbol{\eta}_{\boldsymbol{x}}^T \Sigma_{\boldsymbol{\eta},\boldsymbol{xx}}^{-1} \boldsymbol{\eta}_{\boldsymbol{x}}$ could be found or simulated. In case of normal observations, for instance, it would have a generalized central chi squared distribution. The selection event $\{\boldsymbol{X} = \boldsymbol{x}\}$, however, carries information about the variables in $\boldsymbol{\eta}_{\boldsymbol{x}}$, which is not trivial to formally express.

The selection is decomposed as the intersection of two events $\{\boldsymbol{X} = \boldsymbol{x}\} = \mathrm{X}_1 \cap \mathrm{X}_0$, where X_1 is the event that the variables with label $x_i = 1$ satisfy the selection criterion, and X_0 is the event that the variables with label $x_i = 0$ do not meet the criterion. This decomposition allows to write that $\Sigma_{\boldsymbol{\eta}|\boldsymbol{X},\boldsymbol{xx}} = \mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{X}) = \mathrm{cov}\{(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \mathrm{X}_0) \mid \mathrm{X}_1\}$. As the event X_0 operates on $\boldsymbol{\eta}_{\boldsymbol{x}'}$, the inner conditioning is further decomposed into $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \mathrm{X}_0) = \mathrm{cov}\{E(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}) \mid \mathrm{X}_0\} + E\{\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}) \mid \mathrm{X}_0\}$.

Summarizing the results so far, this section has decomposed the mirror effect into a sequence of conditionings. Under the assumption of symmetry in the vector $\boldsymbol{\beta}$, the decomposition has led to

$$\widehat{m}(n_1) = \frac{1}{n}\sigma^2 \mathrm{tr}\Big(\Sigma_{\boldsymbol{\eta},\boldsymbol{xx}}^{-1}[\mathrm{cov}\{E(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}, \mathrm{X}_0, \mathrm{X}_1)\} + E\{\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}, \mathrm{X}_0, \mathrm{X}_1)\}]\Big) - \frac{n_1}{n}\sigma^2. \quad (23)$$

The conditionings on the events X_0 and X_1 must be made concrete successively and taking into account the precise selection procedure. First, the conditional random vector $\boldsymbol{\eta}_{\boldsymbol{x}'} \mid \mathrm{X}_0$ is considered. From this follow the expected values and covariances for the vectors $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}, \mathrm{X}_0)$ and $E(\boldsymbol{\eta}_{\boldsymbol{x}} \mid \boldsymbol{\eta}_{\boldsymbol{x}'}, \mathrm{X}_0)$, which are functions of the $\boldsymbol{\eta}_{\boldsymbol{x}'} \mid \mathrm{X}_0$. Then the information provided by X_1 is incorporated. Section 4.2 develops the expressions for the case of selection by least angle regression and normally distributed errors.

Normality leads to $\Sigma_{\boldsymbol{\eta}|\mathrm{X}_0,\boldsymbol{xx}} = \Sigma_{\boldsymbol{\eta},\boldsymbol{xx}'}\Sigma_{\boldsymbol{\eta},\boldsymbol{x}'\boldsymbol{x}'}^{-1}\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}'} \mid \mathrm{X}_0)\Sigma_{\boldsymbol{\eta},\boldsymbol{x}'\boldsymbol{x}'}^{-1}\Sigma_{\boldsymbol{\eta},\boldsymbol{xx}'}^T + \Sigma_{\boldsymbol{\eta},\boldsymbol{x}|\boldsymbol{x}'}$, with $\Sigma_{\boldsymbol{\eta},\boldsymbol{x}|\boldsymbol{x}'} = \Sigma_{\boldsymbol{\eta},\boldsymbol{xx}} - \Sigma_{\boldsymbol{\eta},\boldsymbol{xx}'}\Sigma_{\boldsymbol{\eta},\boldsymbol{x}'\boldsymbol{x}'}^{-1}\Sigma_{\boldsymbol{\eta},\boldsymbol{xx}'}^T$ the Schur complement of $\Sigma_{\boldsymbol{\eta},\boldsymbol{xx}}$ in $\Sigma_{\boldsymbol{\eta}}$. In the case where $m < n$, $\Sigma_{\boldsymbol{\eta},\boldsymbol{x}'\boldsymbol{x}'}^{-1}$ denotes the Moore–Penrose generalized inverse of $\Sigma_{\boldsymbol{\eta},\boldsymbol{x}'\boldsymbol{x}'}$.

4.2 The mirror effect in least angle regression

Expression (23) can be evaluated by Monte Carlo simulation. Using a diagonalization made concrete in Assumption 4, this section presents fast, approximate computations that work well in practice. The diagonalized computation of the conditional expectations in (23) is facilitated if the selection events X_0 and X_1 are rewritten in terms of $\boldsymbol{\eta}$, for which Assumption 3 is needed.

The idea is written out below for the case of least angle regression with normal errors. The least angle regression routine (Efron et al., 2004) uses Mallows' C_p as a stopping criterion. The stopping rule implies an optimization in a high-dimensional model, inducing the optimization bias or mirror effect described in this paper. The lasso shrinkage, incorporated in the least angle regression routine or in alternatives such as iterative soft thresholding (Daubechies et al., 2004), compensates for the optimization bias. When the model is used for estimation without shrinkage, the mirror effect must be taken into account in the stopping criterion during variable selection.

Least angle regression selects a variable according to the absolute values of the inner products $\widehat{\boldsymbol{c}} = K^T(\boldsymbol{Y} - K\,\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}}^{\mathrm{LARS}}) = K^T\boldsymbol{\varepsilon} + K^T(K\boldsymbol{\beta} - K\widehat{\boldsymbol{\beta}}_{\boldsymbol{x}}^{\mathrm{LARS}})$. The selection threshold is then $\lambda_{n_1} = |\widehat{c}|_{(n_0:n)}$, this is the n_0th order statistic in vector $\boldsymbol{c}$ of size n, where $n_0 = n - n_1$. The following assumption expresses that the least angle regression routine performs well in identifying the true model.

Assumption 3 *For $n \to \infty$, least angle regression finds a selection $\boldsymbol{x}^*$ of size n_1^* that satisfies two conditions. Firstly, it is sparse, so that $n_1^* = o(n)$. Secondly, it contains the true model except for*

possible small components, so that $\|K^TKE(\widehat{\boldsymbol{\beta}}^{\mathrm{LARS}}_{\boldsymbol{x}^*})-K^TK\boldsymbol{\beta}\|_2^2=\mathcal{O}(n_1^*)$*. In other words, the expected estimator within the selection satisfies the normal equation, up to a bias which is dominated by the estimation variance.*

In any subsequent selection containing $\boldsymbol{x}^*$, the difference $\widehat{\boldsymbol{c}}-\boldsymbol{\eta}=K^TK(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ will primarily depend on the errors, not on the estimation bias. As the variable selection is based on $\widehat{\boldsymbol{c}}$, the ith variable is selected if $|\eta_i|$ is large and the ith component of $K^TK(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is low. The latter term is low if the ith column of K belongs to a multicollinear set of selected columns. Conditioning the selection of a variable on a large value of $|\eta_i|$, we thus have that $q_1(i)=P\left(X_i=1\mid|\eta_i|\geq\lambda_{n_1}\right)$ depends on the relative positions of the columns of K on the n-dimensional unit disk. Assuming a uniform distribution of the columns over the disk, it holds that $q_1(i)\approx n/m$. Similarly, defining $q_0(i)=P\left(X_i=0\mid|\eta_i|<\lambda_{n_1}\right)$, we can write $q_0(i)\approx 1-q_1^*$, where q_1^* is the proportion of nonzeros in $\boldsymbol{\beta}$. In sparse data, $1-q_1^*\approx 1$.

The mirror effect is now computed in several steps, following the expressions of Section 4.1. Let $\boldsymbol{\zeta}_{\boldsymbol{x}'}=V_{\boldsymbol{x}'}^T\boldsymbol{\eta}_{\boldsymbol{x}'}$ be the principal components of the marginal vector $\boldsymbol{\eta}_{\boldsymbol{x}'}$, that is, $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}'})=V_{\boldsymbol{x}'}\Lambda_{\boldsymbol{x}'}V_{\boldsymbol{x}'}^T$, with $\Lambda_{\boldsymbol{x}'}$ a diagonal matrix containing the eigenvalues of the covariance matrix. A similar definition is given for $\boldsymbol{\zeta}_{\boldsymbol{x}}$. Also define $\boldsymbol{d}_{\boldsymbol{x}'}=V_{\boldsymbol{x}'}^T\boldsymbol{c}_{\boldsymbol{x}'}$. Then $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}'}\mid\mathrm{X}_0)=V_{\boldsymbol{x}'}\mathrm{cov}(\boldsymbol{\zeta}_{\boldsymbol{x}'}\mid\mathrm{X}_0)V_{\boldsymbol{x}'}^T$.

Assumption 4 *We assume that conditioning on the* ℓ_∞ *ball* $\mathrm{X}_0=\bigcap_{i|x_i=0}\{\widehat{c}_i^2\leq\lambda_{n_1}^2\}$ *in terms of* $\boldsymbol{c}_{\boldsymbol{x}'}$ *is well approximated by conditioning on the rotated ball* $\mathrm{X}_0^d=\bigcap_{i|x_i^d=0}\{\widehat{d}_i^2\leq\lambda_{n_1}^2\}$*. In this definition, the label* $X_i^d=0$ *means that* $\widehat{d}_i^2<\lambda_{n_1}^2$*. So, we assume that* $\mathrm{cov}(\boldsymbol{\zeta}_{\boldsymbol{x}'}\mid\mathrm{X}_0)\approx\mathrm{cov}(\boldsymbol{\zeta}_{\boldsymbol{x}'}\mid\mathrm{X}_0^d)$.

As the components of $\boldsymbol{\zeta}_{\boldsymbol{x}'}$ are independent, the impact of the event X_0^d can be computed for each component separately, using the result for orthogonal design in (16). Writing $\sigma_i^2=\mathrm{var}(\zeta_{\boldsymbol{x}',i})=\sigma^2\Lambda_{\boldsymbol{x}',ii}$, the statement of (16) reads as $E(\zeta_{\boldsymbol{x}',i}^2\mid\zeta_{\boldsymbol{x}',i}^2>\lambda_{n_1}^2)P(\zeta_{\boldsymbol{x}',i}^2>\lambda_{n_1}^2)=\sigma_i^2P(\zeta_{\boldsymbol{x}',i}^2>\lambda_{n_1}^2)+2\sigma_i^2\lambda_{n_1}\phi_{\sigma_i}(\lambda_{n_1})$. As this expression conditions on the magnitude of $\zeta_{\boldsymbol{x}',i}$, the rules of total probability and Bayes are used to link it to $\{X_i^d=0\}$,

$$\begin{aligned}\mathrm{var}(\zeta_{\boldsymbol{x}',i}\mid X_i^d=0) \;=\; &\Big\{E\left(\zeta_{\boldsymbol{x}',i}^2\mid X_i^d=0,|\zeta_i|<\lambda_{n_1}\right)P\left(X_i^d=0\mid|\zeta_i|<\lambda_{n_1}\right)P\left(|\zeta_i|<\lambda_{n_1}\right)\\&+E\left(\zeta_{\boldsymbol{x}',i}^2\mid X_i^d=0,|\zeta_i|\geq\lambda_{n_1}\right)P\left(X_i^d=0\mid|\zeta_i|\geq\lambda_{n_1}\right)P\left(|\zeta_i|\geq\lambda_{n_1}\right)\Big\}\\&/P(X_i^d=0).\end{aligned}$$

After simplification and introducing $q_1^d=P\left(X_i^d=1\mid|\zeta_i|\geq\lambda_{n_1}\right)\approx n/m$, and $q_0^d=P\left(X_i^d=0\mid|\zeta_i|<\lambda_{n_1}\right)\approx 1$, we get

$$\mathrm{var}(\zeta_{\boldsymbol{x}',i}\mid\mathrm{X}_0^d)\approx\sigma_i^2\left[1-\frac{q_1^d2\lambda_{n_1}\phi_{\sigma_i}(\lambda_{n_1})}{1-q_1^d2\{1-\Phi_{\sigma_i}(\lambda_{n_1})\}}\right]\tag{24}$$

Expression (24) finds the elements of the diagonal covariance matrix $\mathrm{cov}(\boldsymbol{\zeta}_{\boldsymbol{x}'}\mid\mathrm{X}_0^d)$, which approximates $\mathrm{cov}(\boldsymbol{\zeta}_{\boldsymbol{x}'}\mid\mathrm{X}_0)$. Multiplication by $V_{\boldsymbol{x}'}$ leads to the covariance matrix $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}'}\mid\mathrm{X}_0)$, which is used in the computation of $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}}\mid\mathrm{X}_0)$; see Section 4.1. This matrix is then diagonalized as $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}}\mid\mathrm{X}_0)=V_{\boldsymbol{x}}\Lambda_{\boldsymbol{x}}V_{\boldsymbol{x}}^T$, and $\boldsymbol{\zeta}_{\boldsymbol{x}}=V_{\boldsymbol{x}}^T\boldsymbol{\eta}_{\boldsymbol{x}}$. The same type of approximation replaces the event X_1 by a rotated version, leading to $\mathrm{cov}(\boldsymbol{\eta}_{\boldsymbol{x}}\mid\mathrm{X}_0,\mathrm{X}_1)$.

In simulations, the resulting approximate calculation of $\widehat{m}(n_1)$ performs well, meaning that it allows accurate estimation of the prediction error of a least square estimator without shrinkage in a best n_1

	False Positive Percentage			False Negative Percentage			False Discovery Percentage		
	5%	50%	95%	5%	50%	95%	5%	50%	95%
C_p	9.5	16.1	21.9	0	0	0	68.0	75.7	81.2
$C_p + 2\widehat{m}$	0.5	1.8	3.5	0	0	0	10.5	25.0	37.5

Table 2: Quantiles for operating characteristics of least angle regression with and without mirror correction.

term model. Since the approximation assumes that the least angle regression routine reveals the essential terms in the model, problems may occur in cases where this is difficult, in particular, when the number of nonzeros in $\boldsymbol{\beta}$ is large compared to the number of observations n.

4.3 Comparative simulation study

This section investigates the effect on the variable selection of a least angle regression scheme of using the new stopping criterion $C_p(\boldsymbol{X}_{n_1}) + 2\widehat{m}(n_1)$ instead of $C_p(\boldsymbol{X}_{n_1})$. Given the variety of design matrices K and error models, this comparison cannot cover all possible cases. The simulation study generates 200 instances of the model in (1), with a new design matrix K each time, whose elements are all independently chosen from a uniform distribution on $[0, 1]$. The number of observations is $n = 300$, while the number of parameters is $m = 600$. Each parameter β_i is generated independently from a distribution on $\{-1, 0, 1\}$ with probabilities $P(-1) = P(1) = p/2$ and $P(0) = 1 - p$. The sparsity parameter is taken to be $p = 0.05$. The errors are independently, identically distributed $N(0, \sigma^2)$ random variables, so that the signal-to-noise ratio, defined as $\mathrm{SNR} = 10 \log(\|K\boldsymbol{\beta}\|_2^2 / n\sigma^2)$, equals 10.

Table 2 summarizes the empirical values of three operating characteristics. The first is the false positive percentage in each simulation run, defined as 100 times the number of false positives divided by the number of zeros in the parameter vector $\boldsymbol{\beta}$. The second is the false negative percentage, defined as 100 times the number of missed nonzeros divided by the number of nonzeros in the parameter vector. The third is the false discovery percentage, defined as 100 times the number of false positives divided by the number of discoveries. For all three characteristics, the table displays three empirical quantiles.

Both the original C_p criterion and the mirror corrected version find all true nonzeros in $\boldsymbol{\beta}$, there are no false negatives. The original C_p criterion, however, selects much larger models than the mirror corrected criterion, thus containing far more zeros in $\boldsymbol{\beta}$. The median number of zeros selected by the C_p criterion amounts to 16.1% of all the zeros in the full model and to a majority of 75.7% of the selected variables. Measured by the median values of the simulation study, the corrected criterion selects only 1.8% of the zeros, leading to a minority of 25% of false positives among the selected variables. Larger numbers of observations and parameters as well as other design matrices may lead to lower false discovery rates.

Supplementary material

Supplementary Material includes a proof of Proposition 1, the study of the mirror effect for Akaike's information criterion, a few interpretations on the proofs in Appendix A, a discussion on mirror penalties versus penalties proposed in Birgé and Massart (2007), a discussion on Assumption 2, and a note on

accompanying software.

Acknowledgement

Research support by the IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

Appendices

A Proofs of Lemmas 1 and 2

A.1 Proof of Lemma 1

The term $P(S_{n,n_0} \mid \beta = v)$ appearing in the expression (13) for the denominator of (14) can be decomposed into

$$P(S_{n,n_0} \mid \beta_n = v) = \int_{-\infty}^{\infty} P(S_{n,n_0} \mid Y_n = y)\, f_{Y_n|\beta_n}(y \mid v) dy.$$

Given an observation $Y_n = y$, the event S_{n,n_0} occurs if and only if the n_0th order statistic of the remaining $n-1$ observations is above $|Y_n| = |y|$, so $P(S_{n,n_0} \mid Y_n = y) = P(|Y_n|_{(n_0:n-1)} > |y|)$. This is a non-increasing function of $|y|$. Hence

$$\begin{aligned} P(S_{n,n_0} \mid \beta_n = v) &\geq \int_{|y|<|v|} P(S_{n,n_0} \mid Y_n = v)\, f_{Y_n|\beta_n}(y \mid v) dy \\ &= P\left(|Y_n| < |v| \mid \beta_n = v\right) P(S_{n,n_0} \mid Y_n = v). \end{aligned}$$

The second factor can we rewritten as $P(S_{n,n_0} \mid Y_n = v) = P\{U_{(n_0:n-1)} > F_{|Y_n|}(|v|)\}$, where $U_{(n_0:n-1)}$ is the n_0th order statistic of $n-1$ independent uniform variables on $[0,1]$. If $Q_X(p)$ denotes the quantile function of X, and if $v_{\gamma,n} = Q_{|Y_n|}\{Q_{U_{(n_0:n-1)}}(1-\gamma)\}$ for a positive constant γ, then $P(S_{n,n_0} \mid Y_n = v_{\gamma,n}) = \gamma$, and $P(S_{n,n_0} \mid \beta_n = v_{\gamma,n}) \geq P\left(|Y_n| < |v_{\gamma,n}| \mid \beta_n = v_{\gamma,n}\right)\gamma$. Since $n_0/n \to 1$, we can apply Lemma 3, stated below, to arrive at $Q_{U_{(n_0:n-1)}}(1-\gamma) \to 1$ and thus $P(S_{n,n_0} \mid \beta_n = v_{\gamma,n})$ is bounded by $\gamma/2$ in the limit. Moreover, since $P(S_{n,n_0}|\beta_n = v)$ must also be a non-increasing function of $|v|$, the same lower bound holds for any v_n with magnitude below $v_{\gamma,n}$. For all these values, the ratio $\{t(n_1,v_n) - t(n_1,0)\}/r(n_1,v_n)$ thus tends to zero if $\{t(n_1,v_n) - t(n_1,0)\}/v_n^2$ tends to zero.

Among the values of v_n with magnitude below $v_{\gamma,n}$, we first consider the case that v_n is arbitrarily close to 0. Then, for fixed n, we have

$$L_n = \lim_{v\to 0} \frac{t(n_1,v)}{v^2} = \frac{1}{2}\frac{\partial^2 t}{\partial v^2}(n_1,0) = \int_{-\infty}^{\infty} (\sigma^2 - e^2) f_\varepsilon(e) a_n''(e)\, de = \int_{-\infty}^{\infty} \left\{(\sigma^2 - e^2) f_\varepsilon(e)\right\}' a_n'(e)\, de,$$

where $a_n(x) = P(S_{n,n_0} \mid Y_n = x)$. As before, we interpret the event S_{n,n_0} given the observation $Y_n = x$ as $a_n(x) = 1 - F_{U_{(n_0:n-1)}}\left\{F_{|Y_n|}(|x|)\right\}$ The order statistic $U_{(n_0:n-1)}$ of independent, uniform random variables has a Beta distribution with mean $E(U_{(n_0:n-1)}) = n_0/n$ and variance $\mathrm{var}(U_{(n_0:n-1)}) = n_0(n-n_0)/n^2(n+1)$.

Denoting $g(e) = \left\{(\sigma^2 - e^2) f_\varepsilon(e)\right\}'$, we can write

$$L_n = \int_0^\infty \{g(-e) - g(e)\} dF_{U_{(n_0:n-1)}}\left\{F_{|Y_n|}(e)\right\} = E\left[g\{-Q_{|Y_n|}(U_{(n_0:n-1)})\} - g\{Q_{|Y_n|}(U_{(n_0:n-1)})\}\right]$$

We approximate $E\left[g\{Q_{|Y_n|}(U_{(n_0:n-1)})\}\right]$ by $g\{Q_{|Y_n|}(EU_{(n_0:n-1)})\} = g\{Q_{|Y_n|}(n_0/n)\} \to 0$ as $n_0/n \to 1$ when $n \to \infty$. The error of this approximation satisfies

$$\begin{aligned}\left(E\left[g\{Q_{|Y_n|}(U_{(n_0:n-1)})\}\right] - g\{Q_{|Y_n|}(n_0/n)\}\right)^2 &\leq \max_{s\in[0,1]}\left[\frac{dg\{Q_{|Y_n|}(s)\}}{ds}\right]^2 E\left(U_{(n_0:n-1)} - \frac{n_0}{n}\right)^2 \\ &= \max_{u\in\mathbb{R}}\left[\frac{\left\{(\sigma^2 - e^2) f_\varepsilon(e)\right\}''}{f_{|Y_n|}(e)}\right]^2 \mathrm{var}(U_{(n_0:n-1)}).\end{aligned}$$

The factor $\mathrm{var}(U_{(n_0:n-1)}) = \mathcal{O}(n_1/n^2)$ when $n \to \infty$. The factor $\left\{(\sigma^2 - e^2) f_\varepsilon(e)\right\}'' / f_{|Y_n|}(e)$ is bounded for finite u because $f_\varepsilon''(e)$ exists and is finite. It remains bounded for $n \to \infty$ thanks to the sparsity condition in the statement of Lemma 1. The factor tends to zero for infinite u because of the error-free domination condition in Lemma 1, namely $\log f_\varepsilon(e)/\log f_{|Y_n|}(e) \to \infty$. This proves Lemma 1 for v_n arbitrarily close to zero.

For all the other cases, $r(n_1, v_n)$ does not converge to zero, while $t(n_1, 0) \to \int_{-\infty}^{\infty} (\sigma^2 - e^2) f_\varepsilon(e)\, de = 0$. So it suffices to prove that $t(n_1, v_n)/r(n_1, v_n) \to 0$. We pick an arbitrarily small δ and we set $\lambda_{\delta,n} = Q_{|Y_n|}\left\{Q_{U_{(n_0:n-1)}}(1-\delta)\right\}$. Then, $P(S_{n,n_0} \mid Y_n = y) > \delta$ if and only if $|y| < \lambda_{\delta,n}$, hence, for any v,

$$t(n_1, v) = \int_{-\lambda_{\delta,n}-v}^{\lambda_{\delta,n}-v} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y = v + e)\, de + \mathcal{O}(\delta\sigma^2).$$

If $|v_n| < v_{\gamma,n}$, but unbounded by a constant value, then $1/r(n_1, v_n) \sim 1/v_n^2 \to 0$, while $t(n_1, v_n)$ is bounded by σ^2. If $|v_n| < v_{\gamma,n}$ and bounded by a constant, then

$$\lim_{n\to\infty} \int_{-\lambda_{\delta,n}-v_n}^{\lambda_{\delta,n}-v_n} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y = v_n + e)\, de = 0.$$

This follows from Lemma 4, stated below, and from the fact that $\lambda_{\delta,n} \to \infty$ for $n \to \infty$.

Finally, if $|v_n|$ cannot be bounded by $v_{\gamma,n}$ for any positive γ, then for n sufficiently large,

$$\int_{-\lambda_{\delta,n}-v_n}^{\lambda_{\delta,n}-v_n} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y = v_n + e)\, de < \delta,$$

as $\lambda_{\delta,n} - v_n \leq -v_{\gamma,n} \to -\infty$, where we took $\gamma < \delta$. □

The remainder of this section proves the auxiliary lemmas used above.

Lemma 3 *Let U and V be independent and symmetrically distributed around zero and let $f_U(u)$ also be unimodal. Define $W = V + U$. Then, for any value $\alpha \in [0, 1]$, $Q_{|Y|}(\alpha) \geq Q_{|U|}(\alpha)$.*

Proof. It is straightforward to verify that for any value $\alpha \in [0, 1]$, $Q_X(\alpha) \geq Q_Y(\alpha)$ if and only if $F_X(x) \leq F_Y(x)$ for any value $x \in \mathbb{R}$. Second, $F_{|Y|}(x) = F_Y(x) - F_Y(-x)$. We now prove that for positive x it holds that $F_Y(x) \leq F_U(x)$. Similar arguments hold for negative x.

As the distribution of U is symmetric and unimodal, we have for any $x, v > 0$ that $f_U(x+v) \leq f_U(x-v)$, so $F_U(x+v) - F_U(x) \leq F_U(x) - F_U(x-v)$ or $F_U(x+v) + F_U(x-v) \leq 2F_U(x)$.

Then, using the symmetry of V, we can write

$$F_Y(x) = \int_{-\infty}^{\infty} F_U(x-v) f_V(v) dv = E\Big\{F_U(x-V)\Big\} = \frac{1}{2} E\Big\{F_U(x-V) + F_U(x+V)\Big\} \leq F_U(x).$$

□

Corollary 1 *Let $W_n = V_n + U$, where V_n and U are independent and have symmetric distributions. Also suppose that U has a unimodal distribution on $\mathbb{R}$. Then if $\alpha_n \to 1$ for $n \to \infty$, we have $Q_{|W_n|}(\alpha_n) \to \infty$, whatever the distributions of V_n.*

Indeed, $Q_{|W_n|}(\alpha_n) \geq Q_{|U|}(\alpha_n) \to \infty$.

Lemma 4 *Suppose that $0 \leq p_n(x) \leq 1$ is monotone non-decreasing for negative x, and monotone non-increasing for positive x and $\lim_{n\to\infty} p_n(x) = 1$ for any value of x. Also assume that $A = \int_{-\infty}^{\infty} |f(u)| du$ exists and is finite, and define $I_n = \int_{-\lambda_n - c}^{\lambda_n - c} f(u) p_n(u) du$ for constant c. Then, for $\lim_{n\to\infty} \lambda_n = \infty$, $\lim_{n\to\infty} I_n = I = \int_{-\infty}^{\infty} f(u) du$.*

Proof. Consider an arbitrary ε and find a value ℓ^* such that λ_ℓ for $\ell \geq \ell^*$ is sufficiently large in the sense that $\int_{\lambda_\ell - c}^{\infty} |f(u)| du + \int_{-\infty}^{-\lambda_\ell - c} |f(u)| du < \varepsilon$. Then find a value m^* so that for $m > m^* : p_m(\lambda_{\ell^*}) > 1 - \varepsilon$, and define $n^* = \max(\ell^*, m^*)$, then for $n > n^*$,

$$\begin{aligned} |I - I_n| &= \left| \int_{\lambda_n - c}^{\infty} f(u) du + \int_{-\infty}^{-\lambda_n - c} f(u) du + \int_{-\lambda_n - c}^{\lambda_n - c} f(u)\{1 - p_n(u)\} du \right| \\ &\leq \int_{\lambda_\ell - c}^{\infty} |f(u)| du + \int_{-\infty}^{-\lambda_\ell - c} |f(u)| du + \int_{-\lambda_\ell - c}^{\lambda_\ell - c} |f(u)| \{1 - p_n(\lambda_{\ell^*})\} < \varepsilon + A\varepsilon. \end{aligned}$$

□

A.2 Proof of Lemma 2

Until now, we had $m(n_1) \sim t(n_1, 0) = \int_{-\infty}^{\infty} (\sigma^2 - e^2) f_\varepsilon(e) P(S_{n,n_0} \mid Y_n = e)\, de$.

As in the proof of Lemma 1, we use that $P(S_{n,n_0} \mid Y_n = e) = P(|Y_n|_{(n_0:n-1)} > |e|) = 1 - F_{X_n}(|e|)$, where X_n is the n_0th order statistic in n independent observations from $|Y_n|$. Next we define $g_1(e) = (\sigma^2 - e^2) f_\varepsilon(e)$ and we recycle the notation $g(e) = g_1(e) + g_1(-e)$ for different purpose than in the proof of Lemma 1. Finally we introduce $G(e) = \int_0^u g(t) dt$. It holds that $G(0) = 0 = G(\infty)$.

The value of $t(n_1, 0)$ can then be expressed as

$$\begin{aligned} t(n_1, 0) &= \int_{-\infty}^{\infty} g_1(e) \left\{1 - F_{X_n}(|e|)\right\} de = \int_0^{\infty} g(e) \left\{1 - F_{X_n}(e)\right\} de = \int_0^{\infty} G(e) f_{X_n}(e)\, de \\ &= E\left\{G(X_n)\right\} = E\left[G\{Q_{|Y_n|}(U_{(n_0:n-1)})\}\right]. \end{aligned}$$

This proof entails approximating the expression above by $G[Q_{|Y_n|}\{E(U_{(n_0:n-1)})\}] = \int_{-\lambda_n}^{\lambda_n}(\sigma^2-e^2)f_\varepsilon(e) = \tau(\lambda_n)$, where $\lambda_n = Q_{|Y_n|}(EU_{(n_0:n-1)})$. The approximation error satisfies

$$\begin{aligned}\Big(E\Big[G\{Q_{|Y_n|}(U_{(n_0:n-1)})\}\Big] - G\{Q_{|Y_n|}(n_0/n)\}\Big)^2 &\leq \max_{s\in[0,1]}\left[\frac{dG\{Q_{|Y_n|}(s)\}}{ds}\right]^2 E\left(U_{(n_0:n-1)} - \frac{n_0}{n}\right)^2 \\ &= \max_{u\in\mathbb{R}}\left[\frac{(\sigma^2-e^2)f_\varepsilon(e)}{f_{|Y_n|}(e)}\right]^2 \mathrm{var}(U_{(n_0:n-1)}).\end{aligned}$$

Similar arguments as in the proof of Lemma 1 can be used here. The factor $\mathrm{var}(U_{(n_0:n-1)}) = \mathcal{O}(n_1/n^2)$ when $n \to \infty$. The first factor is bounded because of sparsity and dominance assumed in the statement of Lemma 2. Ultimately we find that $|t(n_1, 0) - \tau(\lambda_n)| = \mathcal{O}(n_1^{1/2}/n)$, which is slightly faster than $r(n_1, v_n)$. Indeed, from Expression (8), and taking into account that $E(\varepsilon^2 \mid S_{n,n_0}) \leq \sigma^2$, we easily find that $E\{\mathrm{PE}(\mathcal{J}_{n_1})\} \geq \frac{n_1}{n}\sigma^2$. That lower bound still holds when conditioning on $V_n = v$. □

References

Abramovich, F., Grinshtein, V., and Pensky, P. (2007). On the optimality of Bayesian testimation in the normal means problem. *Annals of Statistics*, 35(5):2261–2286.

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

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138:33–73.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351.

Chen, S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61.

Daubechies, I., Defrise, M., and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57:1413–1457.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59:797–829.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81(3):425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224.

Donoho, D. L. and Johnstone, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 9(1):1–32.

Efron, B., Hastie, T. J., Johnstone, I. M., and Tibshirani, R. J. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499. with discussion.

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, 69(1):131–152.

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical Statistics*, 7(4):469–488.

Ishwaran, H. (2004). Discussion of "Least angle regression" by Efron et al. *Annals of Statistics*, 32(2):452–458.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):1594–1649.

Loubes, J.-M. and Massart, P. (2004). Discussion of "Least angle regression" by Efron et al. *Annals of Statistics*, 32(2):460–465.

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, pages 197–206. University of California Press.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151.

Stine, R. (2004). Discussion of "Least angle regression" by Efron et al. *Annals of Statistics*, 32(2):475–481.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2):1198–1232.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051.

Wainwright, M. J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Woodroofe, M. (1982). On model selection and the arc sine laws. *Annals of Statistics*, 10:1182–1194.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, 93:120–131.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H., Hastie, T. J., and Tibshirani, R. J. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192.

Supplementary Material for Information criteria for variable selection under sparsity

MAARTEN JANSEN
Departments of Mathematics and Computer Science, Université Libre de Bruxelles
B-1050 Brussels, Belgium
maarten.jansen@ulb.ac.be

A Proof of Proposition 1

First we note that the denominator of (19) can be written as

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_\lambda^H) = \frac{1}{n}\sum_{i=1}^n \beta_i^2 P(X_i=0) + \frac{1}{n}\sum_{i=1}^n E\left\{(Y_i-\beta_i)^2 \mid X_i=1\right\} P(X_i=1). \tag{25}$$

We write

$$\kappa(\lambda,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \kappa(\lambda,\beta_i),$$

where

$$\kappa(\lambda,\beta) = E\left\{\mathrm{sign}(\beta+\varepsilon)\,\varepsilon X;\beta\right\} = \int_{-\infty}^{-\lambda-\beta} u f_\varepsilon(u)du + \int_{\lambda-\beta}^{\infty} u f_\varepsilon(u)du,$$

and $X = I(|\beta+\varepsilon|\geq\lambda)$, as before.

We also define the risk contribution of one component, as a function of the threshold value and the component value.

$$r(\lambda,\beta) = \beta^2\, P(|\beta+\varepsilon|<\lambda) + E\left\{\varepsilon^2 \mid |\beta+\varepsilon|\geq\lambda\right\} P(|\beta+\varepsilon|\geq\lambda),$$

so that $R_H(\lambda,\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n r(\lambda,\beta_i)$.

We now establish upper bounds for the one-component relative approximation error

$$\lambda\frac{\kappa(\lambda,\beta)-\kappa(\lambda,0)}{r(\lambda,\beta)},$$

depending on the behavior of β as λ increases. The threshold increases as function of n. The subsequent analysis shows that, whatever the behavior of a particular component with increasing n, the one-component relative approximation error tends to zero. The dependence from n in threshold and component is omitted in the subsequent notations.

We distiguish between $|\beta|-\lambda$ bounded from above and the case where $|\beta|-\lambda$ is positive and unbounded.

First we consider the case that $-\lambda-\Gamma\leq\beta\leq\lambda+\Gamma$, with Γ an arbitrary real number. We start from the lower bound, valid in any case,

$$r(\lambda,\beta) \geq \sigma_0^2\, P(|\beta+\varepsilon|>\lambda) + \beta^2\, P(|\beta+\varepsilon|\leq\lambda) \geq \beta^2\, P(|\beta+\varepsilon|\leq\lambda).$$

Furthermore we have for $0 \leq \beta \leq \lambda + \Gamma$ that the factor

$$\begin{aligned} P(|\beta + \varepsilon| \leq \lambda) &= P(\beta + \varepsilon \leq \lambda) - P(\beta + \varepsilon \leq -\lambda) = P(\varepsilon \leq \lambda - \beta) - P(\varepsilon \leq -\lambda - \beta) \\ &\geq P(\varepsilon \geq \Gamma) - P(\varepsilon \geq \lambda), \end{aligned}$$

and the same expression holds for $-\lambda - \Gamma \leq \beta \leq 0$, with a similar proof.

This allows us to concentrate on the function

$$\gamma(\lambda, \beta) = \frac{\lambda}{\beta^2} \Big\{ \kappa(\lambda, \beta) - \kappa(\lambda, 0) \Big\}.$$

We prove the following lemma:

Lemma 5 *If $E(\varepsilon^{2+\rho})$ exists and is finite for some positive ρ and the density function $f_\varepsilon(u)$ is symmetric and has a converging derivative for $u \to \pm\infty$, then the function $\gamma(\lambda, \beta)$, defined above, satisfies*

$$\lim_{\lambda \to \infty} \gamma\{\lambda, \beta(\lambda)\} = 0,$$

for any function $\beta(\lambda)$ bounded by $\pm(\lambda + \Gamma)$, where Γ is zero or a positive real number.

Proof. Consider an arbitrarily small $\delta > 0$. We will prove that there exists a value λ^* so that if $\lambda > \lambda^*$ it holds that $\gamma\{\lambda, \beta(\lambda)\} < \delta$.

We first consider the case that $\beta(\lambda)$ for some λ is arbitrarily close to zero. It is easy to verify that for fixed λ, and symmetric $f_\varepsilon(u)$,

$$\lim_{\beta \to 0} \gamma(\lambda, \beta) = \frac{1}{2} \lambda \frac{\partial^2 \kappa}{\partial \beta^2}(\lambda, 0) = -\lambda \Big\{ \lambda f_\varepsilon(\lambda) \Big\}'.$$

Then for any positive λ, there exists a value β_0, so that for any β with $|\beta| < \beta_0$,

$$|\gamma(\lambda, \beta)| < \lambda \left| \{\lambda f_\varepsilon(\lambda)\}' \right| + \delta/2.$$

Moreover, as $E(\varepsilon^2)$ is finite, we have $\lim_{u \to \infty} u^2 f_\varepsilon(u) = 0$. Since $\lim_{u \to \infty} f'_\varepsilon(u)$ exists, which for a density function means it must be zero, we can apply de l'Hôpital's rule to find

$$0 = \lim_{u \to \infty} \frac{u f_\varepsilon(u)}{1/u} = \lim_{u \to \infty} \frac{\{u f_\varepsilon(u)\}'}{-1/u^2} = \lim_{u \to \infty} u \; u\{u f_\varepsilon(u)\}'$$

and thus

$$\lim_{u \to \infty} u\{u f_\varepsilon(u)\}' = 0.$$

Hence, there exists a λ_1^* such that for $\lambda > \lambda_1^*$,

$$\lambda \left| \{\lambda f_\varepsilon(\lambda)\}' \right| < \delta/2.$$

We thus have found a value β_0, independent from λ, for which

$$|\gamma(\lambda, \beta)| < \delta,$$

if $|\beta| < \beta_0$ and λ sufficiently large.

Second, we consider the case that $\beta(\lambda)$ for a given value λ is small, but not arbitrarily close to zero. More precisely, suppose that $\beta_0 < |\beta| < \lambda - \lambda^{1/(1+\rho)}$, with $0 < \rho$, then

$$|\gamma(\lambda, \beta)| \leq \frac{\lambda}{\beta_0^2} \int_{\lambda^{1/(1+\rho)}}^{\infty} u f_\varepsilon(u)\, du,$$

and

$$\begin{aligned} \lim_{\lambda\to\infty} \lambda \int_{\lambda^{1/(1+\rho)}}^{\infty} u f_\varepsilon(u)\, du &= \lim_{x\to\infty} \frac{\int_x^\infty u f_\varepsilon(u)\, du}{x^{-(1+\rho)}} \\ &= \lim_{x\to\infty} \frac{x f_\varepsilon(x)}{(1+\rho)x^{-(2+\rho)}} \\ &= \frac{1}{1+\rho} \lim_{x\to\infty} x^{3+\rho} f_\varepsilon(x). \end{aligned}$$

The latter limit must be zero in order for $E(\varepsilon^{2+\rho})$ to be finite. We thus have a value λ_2^* above which $|\gamma(\lambda, \beta)| \leq \delta$ if $\beta_0 < |\beta| < \lambda - \lambda^{1/(1+\rho)}$.

Third, we concentrate on values β close to the threshold value λ, namely $\lambda - \lambda^{1/(1+\rho)} < |\beta| < \lambda + \Gamma$. As

$$|\kappa(\lambda, \beta)| \leq E(|U|)/2,$$

we can write, for $\lambda \to \infty$,

$$|\gamma(\lambda, \beta)| \leq \frac{\lambda}{\left(\lambda - \lambda^{1/(1+\rho)}\right)^2} E(|U|) \to 0,$$

leading to the conclusion that there exists a value λ_3^* above which $\Big|\gamma(\lambda, \beta)\Big| \leq \delta$ if $\lambda - \lambda^{1/(1+\rho)} < |\beta| < \lambda$. Taking $\lambda^* = \max_{i=1,2,3} \lambda_i^*$ concludes the proof of Lemma 5. □

In order to finish the proof of Proposition 1, we have to consider one more case, that of unbounded $\zeta(\lambda) = |\beta| - \lambda$.

If $\zeta(\lambda)$ grows at least as λ^ρ with postive ρ, possibly smaller than 1, then the exponential decay of $f_\varepsilon(u)$ ensures that $\lambda\,[\kappa(\lambda, \beta) - \kappa(\lambda, 0)] \to 0$, while $r(\lambda, \beta)$ converges to σ^2.

Otherwise, that is, if $\zeta(\lambda)/\lambda \to 0$, then the dominating term in $\lambda\,\kappa(\lambda, \beta)$ is

$$\lambda\,\kappa(\lambda, \beta) \sim \lambda \int_{-\zeta(\lambda)}^{\infty} u f_\varepsilon(u) du.$$

The contribution to the prediction error is bounded from below by

$$r(\lambda, \beta) \geq \Big\{\lambda + \zeta(\lambda)\Big\}^2 \int_{-2\lambda-\zeta(\lambda)}^{-\zeta(\lambda)} f_\varepsilon(u) du.$$

The ratio of these two converges to zero (as follows from applying de L'Hôpital's rule).

All together, we conclude that

$$\frac{\lambda\,|\kappa(\lambda, \boldsymbol{\beta}) - \kappa(\lambda, 0)|}{\sigma_0^2 \frac{EN_1}{n} + \frac{1}{n}\sum_{i=1}^n \beta_i^2 P(X_i = 0)} \leq \max_{\beta\in\mathbb{R}} \frac{\lambda\,|\kappa(\lambda, \beta) - \kappa(\lambda, 0)|}{r(\lambda, \beta)} \to 0,$$

for $\lambda \to \infty$.

B The Mirror effect for Akaike's Information Criterion

Akaike's Information Criterion estimates the Kullback–Leibler distance of a model with respect to true, unobserved distributions of the observations. Let $g_{\boldsymbol{Y}}(\boldsymbol{y})$ be the joint density of n independent observations $\boldsymbol{Y}$, and let $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})$ be a model for these observations, with p parameters in $\boldsymbol{\theta}$, then the Kullback–Leibler distance equals

$$\mathrm{KL}\{g, f(\cdot,\boldsymbol{\theta})\} = \frac{1}{n}\sum_{i=1}^{n}\left\{E_g \log g_i(Y_i) - E_g \log f_i(Y_i;\boldsymbol{\theta})\right\}, \tag{26}$$

where g_i and f_i are the true and model marginal densities. It is obvious that the terms $E_g \log g_i(Y_i)$ acts as constants, and so model selection concentrates on the sum

$$H_{\boldsymbol{x}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} E_g\{\log f_i(Y_i;\boldsymbol{\theta})\}. \tag{27}$$

In this notation $H_{\boldsymbol{x}}$, the subscript $\boldsymbol{x}$ refers to the model under consideration. As introduced in Section 2, $\boldsymbol{x}$ is a binary vector of length n where the ones correspond to the parameters that are estimated in the model, whereas the zeros are parameters that are not included in this particular model.

At this point, we restrict discussion to independent, homoscedastic, normally distributed data, that is, the true model can be written as $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a zero mean normal vector with constant variance σ^2. This true model belongs to the space of models considered in our selection procedure. Let $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}^2$ in a model $\boldsymbol{x}$ be values of the unknown parameters under consideration, then

$$\begin{aligned} H_{\boldsymbol{x}}(\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2) &= \frac{1}{n}\sum_{i=1}^{n} E_g\left\{-\frac{(Y_i-\widetilde{\beta}_i)^2}{2\widetilde{\sigma}^2} - \frac{1}{2}\log(2\pi\widetilde{\sigma}^2)\right\} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\beta_i-\widetilde{\beta}_i)^2+\sigma^2}{2\widetilde{\sigma}^2} + \frac{1}{2}\log(2\pi\widetilde{\sigma}^2)\right\}. \end{aligned}$$

In practice, the values $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}^2$ follow from an estimation procedure within a selected model. As a consequence, the outcome is random, say $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$, and hence, so is the value of $H_{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2)$.

Since we cannot evaluate $H_{\boldsymbol{x}}(\widetilde{\boldsymbol{\beta}},\widetilde{\sigma}^2)$ because of the unobserved $\boldsymbol{\beta}$ and σ^2, we substitute the expected value operator E_g by its empirical counterpart, based on an estimator of the unknown parameters, thus defining

$$\widehat{Q}_{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}},\widehat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{(Y_i-\widehat{\beta}_i)^2}{2\widehat{\sigma}^2} - \frac{1}{2}\log(2\pi\widehat{\sigma}^2)\right\}. \tag{28}$$

Imposing a variance estimator based on the residuals, $\widehat{\sigma}^2 = n_0^{-1}\sum_{i=1}^{n}(Y_i-\widehat{\beta}_i)^2$, we arrive at

$$\widehat{Q}_{\boldsymbol{x}} = -\frac{1}{2}\frac{n_0}{n} - \frac{1}{2}\log(2\pi\widehat{\sigma}^2).$$

In this expression, $n_0 = n - n_1$, where n_1 is number of nonzeros in the model $\boldsymbol{x}$.

The difference in expectation between $H_{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and $\widehat{Q}_{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ equals

$$E(\widehat{Q}_{\boldsymbol{x}} - H_{\boldsymbol{x}}) = -\frac{1}{2}\frac{n_0}{n} + \frac{1}{2}E\left[\frac{\sigma^2}{\widehat{\sigma}^2}\,\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(\beta_i - \widehat{\beta}_i)^2}{\sigma^2} + 1\right\}\right]. \tag{29}$$

In the case of component selection in a sparsity model, that is, $\widehat{\beta}_i = Y_i\, x_i$, with $\boldsymbol{x}$ the (not yet random) model under consideration, we have that $\widehat{\sigma}^2 = n_0^{-1}\sum_{i\in\mathcal{I}_0} Y_i^2$ and $\sum_{i=1}^{n}(\beta_i - \widehat{\beta}_i)^2 = \sum_{i\in\mathcal{I}_0}\beta_i^2 + \sum_{i\in\mathcal{I}_1}(Y_i - \beta_i)^2$, where $\mathcal{I}_1$ is the set of indices corresponding to the ones in vector $\boldsymbol{x}$ and $\mathcal{I}_0$ the complementary set. As $\mathcal{I}_1$ and $\mathcal{I}_0$ are disjoint sets, both factors in the product of (29) have no common random term, so these factors are independent. Moreover, under the assumption that $\beta_i = 0$ if $i \in \mathcal{I}_0$, we have $n_0\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{n_0}$, so $E(\sigma^2/\widehat{\sigma}^2) = n_0/(n_0 - 2)$. All this leads to

$$E(\widehat{Q}_{\boldsymbol{x}} - H_{\boldsymbol{x}}) = -\frac{1}{2}\frac{n_0}{n} + \frac{1}{2}\left(\frac{n_0}{n_0 - 2}\,\frac{n + n_1}{n}\right) = \frac{n_0(n_1 + 1)}{n(n_0 - 2)} \sim \frac{n_1 + 1}{n}.$$

Defining Akaike's Information Criterion as

$$\mathrm{AIC}(\boldsymbol{x}) = 2\widehat{Q}_{\boldsymbol{x}} - 2\frac{n_1 + 1}{n}, \tag{30}$$

we see that $E\{\mathrm{AIC}(\boldsymbol{x})\} \approx 2EH_{\boldsymbol{x}}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$, where $\widehat{\beta}_i = Y_i\, x_i$ and $\widehat{\sigma}^2 = n_0^{-1}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_i)^2$.

If $\boldsymbol{X}$ is found by minimization of $E\{\mathrm{AIC}(\boldsymbol{x})\}$ for given n_1, then $\mathcal{I}_0$ is no longer a fixed, but a random set and at the same time, the zero mean components in this set are no longer independently, normally distributed. The two factors in (29), conditionally independent on $\boldsymbol{X}$, are now dependent.

In sparsity models, the n_1 parameters of the selected model are the positions of the nonzero elements in $\widehat{\boldsymbol{\beta}}$. The optimal value of $\mathrm{AIC}(\boldsymbol{x})$ for given n_1 is obtained by choosing the n_1 observations in $\boldsymbol{Y}$ with largest magnitude. As in Section 2.3, $\mathcal{X}_{n_1}$ stands for the random set of selected components and we let $\widehat{\sigma}^2_{n_1} = n_0^{-1}\sum_{i\in\mathcal{X}'_{n_1}} Y_i^2$. We further denote K_{n_1} and $\widehat{Q}_{n_1}$ for the values of $H_{\boldsymbol{x}}$ and $\widehat{Q}_{\boldsymbol{x}}$ corresponding to $\mathcal{X}_{n_1}$. We can write

$$E(\widehat{Q}_{n_1} - K_{n_1}) = -\frac{1}{2}\frac{n_0}{n} + \frac{1}{2}E\left[\frac{\sigma^2}{\widehat{\sigma}^2_{n_1}}\left\{1 + \frac{1}{n}\sum_{i\in\mathcal{X}_{n_1}}\frac{\varepsilon_i^2}{\sigma^2} + \frac{1}{n}\sum_{i\in\mathcal{X}'_{n_1}}\frac{\beta_i^2}{\sigma^2}\right\}\right].$$

We have that

$$\frac{1}{n}\sum_{i\in\mathcal{X}_{n_1}}\varepsilon_i^2 \xrightarrow{P} \sigma^2 - \frac{1}{n}\sum_{i\in\mathcal{X}'_{n_1}}\varepsilon_i^2.$$

Again considering an independent, identically distributed random model for $\boldsymbol{\beta}$, as in Section 2.3, we find

$$\frac{1}{n}\sum_{i\in\mathcal{X}_{n_1}}\varepsilon_i^2 \xrightarrow{P} \sigma^2 - \frac{n_0}{n}E(\varepsilon^2 \mid A_{n,n_0}).$$

For the penalty in Akaike's information criterion after selection, this becomes

$$E(\widehat{Q}_{n_1} - K_{n_1}) \to -\frac{1}{2}\frac{n_0}{n} + \frac{1}{2}\left\{\frac{\sigma^2 + \sigma^2 - \frac{n_0}{n}E(\varepsilon^2 \mid A_{n,n_0}) + \frac{n_0}{n}E(\beta_n^2 \mid A_{n,n_0})}{E(Y_n^2 \mid A_{n,n_0})}\right\}.$$

As in Section 2.3, Assumption 2 and its implication (9), we can omit $o(n_1/n)$ terms to arrive at

$$E(\widehat{Q}_{n_1} - K_{n_1}) \sim -\frac{1}{2}\frac{n_0}{n} + \frac{1}{2}\left\{\frac{2\sigma^2 - \frac{n_0}{n}E(\varepsilon^2 \mid A_{n,n_0})}{E(\varepsilon^2 \mid A_{n,n_0})}\right\} = \frac{\sigma^2}{\sigma_A^2} - \frac{n_0}{n},$$

where $\sigma_A^2 = E(\varepsilon^2 \mid A_{n,n_0})$.

If we denote $\sigma_\lambda^2 = E(\varepsilon^2 \mid -\lambda_n < Y_n < \lambda_n)$ and $\widetilde{n}_0 = n\,P(|U| < \lambda_n)$, then for λ_n defined in Lemma 2 of the main article, this Lemma states that

$$\frac{n_0}{n}(\sigma^2 - \sigma_A^2) \sim \frac{\widetilde{n}_0}{n}(\sigma^2 - \sigma_\lambda^2),$$

where the asymptotic equivalence is relative with respect to the risk; see the Lemma for details. We can write

$$E(\widehat{Q}_{n_1} - K_{n_1}) \sim \frac{\sigma^2}{\sigma_A^2} - \frac{n_0}{n} = \frac{\sigma^2}{\sigma_A^2}\frac{n_1}{n} + \frac{1}{\sigma_A^2}\frac{n_0}{n}(\sigma^2 - \sigma_A^2) \sim \frac{\sigma^2}{\sigma_A^2}\frac{n_1}{n} + \frac{1}{\sigma_A^2}\frac{\widetilde{n}_0}{n}(\sigma^2 - \sigma_\lambda^2).$$

Furthermore, we have

$$\sigma_A^2 \sim \sigma_\lambda^2 + \left(1 - \frac{\widetilde{n}_0}{n_0}\right)(\sigma^2 - \sigma_\lambda^2) = \sigma_\lambda^2 + \mathcal{O}\left(\frac{n_1}{n - n_1}\right),$$

so we can replace σ_A^2 by σ_λ^2, leading to

$$\mathrm{AIC}^m(n_1) = 2\widehat{Q}_{n_1} - 2E(\widehat{Q}_{n_1} - K_{n_1}) = -\frac{n_0}{n} - \log(2\pi\widehat{\sigma}^2) - 2\frac{n_1}{n}\frac{\sigma^2}{\sigma_\lambda^2} - 2\frac{\widetilde{n}_0}{n}\left(\frac{\sigma^2}{\sigma_\lambda^2} - 1\right), \quad (31)$$

which is (21) in the article.

The mirror effect in Akaike's criterion is illustrated in Figure 2. The setup for the simulation displayed in Figure 2 has been the same as that of Figure 1 in the main article.

C Remarks about the proofs in Appendix A

Remark 1 *The convergence analyses of both approximating expressions for the mirror rely on upper-bounds for expressions of the form*

$$\max_{u\in\mathbb{R}}\left\{\frac{(\sigma^2 - e^2)f_\varepsilon(e)}{f_{|Y_n|}(e)}\right\}^2 \mathrm{var}(U_{(n_0:n-1)}).$$

We have found that the second factor converges, but just a bit faster than $E\mathrm{PE}(\mathcal{J}_{n_1})$*. In practice, however, the first factor converges as well. Indeed, instead of taking the maximum over all* $u \in \mathbb{R}$*, we can consider* u *in the neighborhood of* $\lambda_n = Q_{|Y_n|}(n_0/n)$*, which tends to infinity. The heavier tail of the error-free distribution then induces faster convergence. In the case of normal errors and a Laplace prior for the noise-free data, for instance, additional convergence rate is of the order* $\mathcal{O}\left(\exp[\{\log(n)\}^{1/2}]/n\right)$*, which is just a little slower than* $\mathcal{O}\{\log(n)/n\}$*.*

Remark 2 *The analyses of the approximating mirror expressions rely on the exact Beta distribution of* $U_{(n_0:n-1)}$*, necessary knowledge in the elaboration of its variance. This exact calculation, however, is based on the assumption that the observations, and so the errors, are mutually independent. Nevertheless, it can be conjectured that even for dependent or correlated errors, the approximating expression for the mirror effect still holds true.*

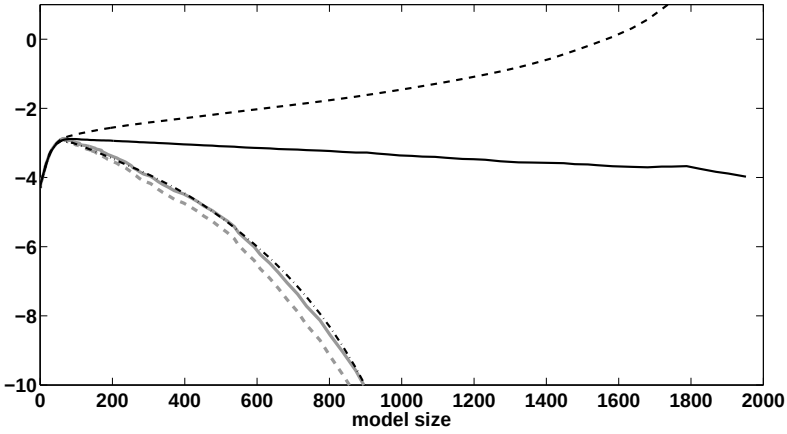


Figure 2: Akaike's information criterion for best n_1 term selection with mirror effect. The solid line is minus the logarithm of the Kullback–Leibler distance for the first n_1 variables of a sequence arranged by an oracle that placed all observations in descending order of magnitude of β_i. The dashed, increasing line is Akaike's information criterion in its classical form applied to the n_1 largest observations in $\boldsymbol{Y}$. This curve cannot be used to locate the correct extremum of the Kullback–Leibler curve. Its reflection with respect to the oracle curve coincides approximately with the dot-dashed line, which is minus the logarithm of the Kullback–Leibler distance for selection of the n_1 largest observations in $\boldsymbol{Y}$. This curve is well estimated by the mirror corrected expressions for Akaike's information criterion, depicted in grey colors, stated in Equation (21) of the main article. When the variance is estimated using generalized cross validation (solid grey line), the outcome is better, compared to a variance estimation using median absolute deviation.

D The mirror penalty and the penalties of Birgé and Massart

The mirror correction can be seen as a modification of the penalty in a variable selection criterion, taking the sparsity into account.

In the case of Mallows's C_p criterion, the mirror corrected version can be written as

$$\widetilde{\Delta}_p(n_1) = \frac{1}{n}\|\widehat{\boldsymbol{y}} - \boldsymbol{Y}\|_2^2 + 2\frac{n_1}{n}\sigma^2 + 2\frac{n_0}{n}\Big\{\sigma^2 - E(\varepsilon^2|A_{n,n_0})\Big\},$$

where A_{n,n_0} is the event that Y_n is among the n_0 smallest observations in a sample of size n, and $n_1 = n - n_0$ is the number of non-small observations, i.e., the size of the selected set of variables. In a sense explained in the paper, the criterion can be approximated by the expression

$$\widetilde{\Delta}_p(n_1) \approx \frac{1}{n}\|\widehat{\boldsymbol{y}} - \boldsymbol{Y}\|_2^2 + 2\frac{n_1}{n}\sigma^2 + 2\sigma^2\int_{-\lambda_{n_1}}^{\lambda_{n_1}}\left(1 - \frac{u^2}{\sigma^2}\right)f_\varepsilon(u)du,$$

where $f_\varepsilon(u)$ is the error density and $\lambda_{n_1} = Q_{|Y|}(1 - n_1/n)$, with $Q_{|Y|}(\alpha)$ the quantile function of the magnitude of the observation Y in a Bayesian model $Y = \beta + \varepsilon$. The approximative criterion reduces, in the case of normal errors, to

$$\widetilde{\Delta}_p(n_1) \approx \frac{1}{n}\|\widehat{\boldsymbol{y}} - \boldsymbol{Y}\|_2^2 + 2\frac{n_1}{n}\sigma^2 + 4\sigma_2\lambda_{n_1}\phi_\sigma(\lambda_{n_1}),$$

where $\phi_\sigma(x)$ is the normal probability density function with zero mean and variance σ^2.

An important benchmark in this respect is the minimum penalty resulting from the analysis by Birgé and Massart (2007).

Before comparing the mirror penalty with the minimum penalty, I first list the main differences in approach and results between their and my paper.

1. The newly proposed viewpoint of the problem as a mirror effect allows to establish a sparsity correction for selection criteria other than Mallows's C_p, the case of AIC being worked out in the text. The mirror correction is also possible for error densities other than normal.

2. The result of the new analysis is not a lower bound on the penalty, but a data-dependent penalty. The data-dependency is realized by a threshold value λ_{n_1} which is a quantile of the observations Y in a Bayesian model. The Bayesian description has no further impact in the practical implementation if we estimate λ_{n_1} by its empirical counterpart. The threshold appears in the bounds of an integration of a function depending only on the error distribution. The threshold thus expresses exactly what the mirror effect is about: given the number of selected variables n_1, the threshold corresponding to n_1 is a matter of the interaction between signal and noise, but once λ_{n_1} has been set, the correction necessary for its quality assessment is a matter of false positives created by error effects only.

Birgé and Massart present a lower bound that avoids inconsistent estimators, although penalties below the bound do not necessarily lead to problems (Birgé and Massart, 2007, page 42). The presented lower bound is of the form

$$\mathrm{pen}_{\mathrm{BM}}(x) = Kx_1\sigma^2\left[1 + 2\log(1/x_1) + 2\{\log(1/x_1)\}^{1/2}\right],$$

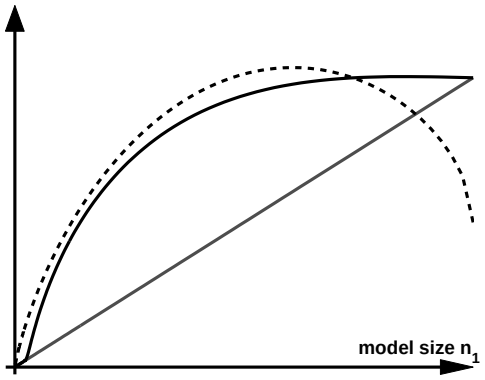


Figure 3: Mirror penalty versus Birgé-Massart lower bound. In grey line: Mallows's penalty term. In solid black line: the mirror corrected penalty. In dashed line: the lower bound proposed by Birgé and Massart. The mirror penalty is adaptive to the sparsity in the data. The steep increase is therefore deferred to model sizes where errors start to play a role in the selection process.

where $x_1 = n_1/n$. The mirror penalty, presented in the paper, can be written as

$$\mathrm{pen}_{\mathrm{mir}}(x_1) = 2x_1\sigma^2 + 4\sigma^2\lambda_{n_1}\phi_\sigma(\lambda_{n_1}).$$

Figure 3 compares the mirror penalty with an implementation of the lower bound for a typical case, further explained below. We can draw the following conclusions.

1. Although at first sight, it seems that the mirror penalty violates the lower bound, the lower bound should not be checked for its absolute value, but rather for its slope. Indeed, while a constant may be added to all possible models to ensure that penalties are above a minimum, a steep slope discourages models with too many selected variables.

2. The figure illustrates the adaptive nature of the mirror penalty: small models include only highly significant variables. In the selection of those, there is no need to take any error effect into account. In that range, the distinction between significant variables and the errors is so clear that the non-linear selection acts as an oracle that knows the order of the error-free values of $\boldsymbol{\beta}$. Such an oracle can rely on Mallows's penalty as a stopping criterium in selecting the right number of variables. From a certain value of n_1, depending on the signal at hand, the errors play a role in the selection procedure, resulting in a steep slope of the penalty in order to keep these effects under control. Birgé's and Massart's lower bound is not data-dependent, which explains the steep slope from the beginning.

In order to verify that the mirror penalty increases sufficiently fast as soon as observational errors affect the selection, we first consider the case where the observations contain only errors and no signal (i.e., $\beta_i = 0$). Let $\lambda_{n_1}^{(0)} = Q_{|\varepsilon|}(1 - n_1/n) = \Phi_\sigma^{-1}(1 - x_1)$, then the penalty

$$\mathrm{pen}_{\mathrm{mir0}}(x_1) = 2x_1\sigma^2 + 4\sigma^2\lambda_{n_1}^{(0)}\phi_\sigma(\lambda_{n_1}^{(0)}).$$

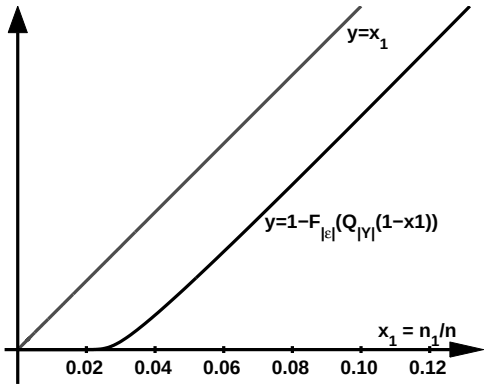


Figure 4: Plot of function $y(x_1) = 1 - F_{|\varepsilon|}\{Q_{|Y|}(1 - x_1)\}$ near origin. This function connects a data-independent version of the mirror penalty to the actual, data-adaptive mirror penalty.

does not depend on $F_{|Y|}$. Moreover, it can be verified that

$$\frac{d\mathrm{pen}_{\mathrm{mir0}}(x_1)}{dx_1} \geq \frac{d\mathrm{pen}_{\mathrm{BM}}(x_1)}{dx_1}$$

for x_1 near 0.

Next, define $y(x_1) = 1 - F_{|\varepsilon|}\{Q_{|Y|}(1 - x_1)\}$, then $\mathrm{pen}_{\mathrm{mir}}(x_1) = \mathrm{pen}_{\mathrm{mir0}}\{y(x_1)\}$. The function $y(x_1)$ is a bijection on $[0, 1]$, whose behavior near 0 is depicted in Figure 4 for the same model as in Figure 3. In these figures. the error-free data are modelled as zero inflated double exponential variables, i.e., $f_\beta(\beta \mid \beta \neq 0) = (a/2)\exp(-a|\beta|)$, where in the figures the hyperparameter values $a = 1/5$ and $p = P(\beta \mid \beta \neq 0) = 0.05$ were used. The model allows to elaborate analytically or numerically all expected values without any simulation. By definition, it holds that $y(1) = 1$, while $y(x_1) \leq x_1$. The function $y(x_1)$ is thus a bijective shrinking function. As a consequence, the behavior of $\mathrm{pen}_{\mathrm{mir0}}(x_1)$ is inherited by the function $\mathrm{pen}_{\mathrm{mir}}(x_1) = \mathrm{pen}_{\mathrm{mir0}}(y(x_1))$, but with some delay. This implies that $\mathrm{pen}_{\mathrm{mir}}(x_1)$ shows a steep increase as soon as error effects appear in the selection process.

E The interpretation of Assumption 2

Assumption 2 is expressed within the setting of a random model for the error-free parameters β_n as $E(\beta_n^2 \mid S_{n,n_0}) = o(n_1/n)$. Translated into a fixed parameter model with a vector $\boldsymbol{\beta}_n = (\beta_{n,1}, \beta_{n,2}, \ldots, \beta_{n,n})$, this becomes an expected average over all not selected i:

$$\frac{1}{n_0} E\left(\sum_{i \in \mathcal{X}'_{n_1}} \beta_{n,i}^2\right) = o(n_1/n).$$

This can be rewritten as

$$\frac{1}{n_0}E\left(\sum_{i\in\mathcal{X}'_{n_1}}\beta^2_{n,i}\right)=\frac{1}{n_0}E\left(\sum_{i=1}^{n}\beta^2_{n,i}I(i\in\mathcal{X}'_{n_1})\right)=\frac{n}{n_0}\frac{1}{n}\sum_{i=1}^{n}\beta^2_{n,i}P(i\in\mathcal{X}'_{n_1}),$$

leading to the formulation $n^{-1}\sum_{i=1}^{n}\beta_i^2P(|Y_i|<\lambda)=o(n_1n^{-1})$, as found in the article, right after the statement of Assumption 2.

The assumption can be interpreted as a bound on the lost information due to false negatives or missed discoveries. It imposes a three-fold condition:

1. the vector $\boldsymbol{\beta}_n$ is sparse;
2. the errors do not hinder a good separation between significant and insignificant components in $\boldsymbol{\beta}_n$. More precisely, the tail of the error distribution is not heavy, excluding large noise components that could interfere with significant components in $\boldsymbol{\beta}_n$;
3. the model size n_1, or the threshold, is well chosen by the variable selection algorithm, so that it indeed separates between significant and insignificant components.

The remainder of this section discusses the three conditions in a quantitative way.

The notion of sparsity is defined in an asymptotic way, imposing that for $n\to\infty$, the vector $\boldsymbol{\beta}_n$ becomes sparser, while its mean squared value is assumed to be constant. This can be formalized by defining an invertible, non-decreasing, positive function $\beta_n(x)$, defined on $[0,1]$ so that the ordered absolute vector components satisfy $|\beta|_{n,(i)}=\beta_n(i/n)$. Sparsity means that $\|\beta_n\|_2^2=1$, while for some $p<2$, $\beta_n(x)\in L_p(r_n)$ with $r_n\to 0$. The L_p ball with radius r_n contains all functions β for which $\|\beta_n\|_p\le r_n$, where $\|\beta_n\|_p=\int_0^1\beta_n^p(x)dx$, for $0<p\le 2$.

We define an index of sparsity $x_1(n)\in[0,1]$ as the value for which

$$\int_0^{1-x_1(n)}\beta_n^2(x)dx=x_1(n). \tag{32}$$

This L_2-concentration index can be seen as the equivalent of the g-index in bibliometry (Egghe, 2006), where sparsity corresponds to a low index value. If $x_1(n)$ is small, then the greater part $(1-x_1(n))$ of the energy in the vector $\boldsymbol{\beta}_n$ is concentrated the large components, accounting for only a fraction $x_1(n)$ of the total size of the vector. This concentration is guaranteed for functions in L_p balls, as follows from the next lemma.

Lemma 6 *If $\beta_n(x)\in L_p(r_n)$, then $x_1(n)\le r_n^{2p/(2-p)}\{1-x_1(n)\}$.*

Searching for a variable selection satisfying $\sum_{i=1}^{n}\beta^2_{n,i}P(i\in\mathcal{X}'_{n_1})=o(n_1)$, we look for model sizes n_1 close to $nx_1(n)$.

For $n_1=nx_1(n)$, and denoting $\widetilde{x}_1(n)=\int_{1-x_1(n)}^1\beta_n^2(x)\,P\{\widehat{\beta}=0\mid\beta=\beta_n(x)\}\,dx$, we find

$$\int_0^1\beta_n^2(x)P\{\widehat{\beta}=0\mid\beta=\beta_n(x)\}dx\le\int_0^{1-x_1(n)}\beta_n^2(x)\,dx+\widetilde{x}_1(n)=x_1(n)+\widetilde{x}_1(n).$$

Neglecting the small probability $P\{Y_i < -\lambda_n \mid \beta = \beta_n(x)\}$, for $\lambda_n = \beta_n\{1 - x_1(n)\}$, the second term can be bounded by

$$\widetilde{x}_1(n) \leq \int_{1-x_1(n)}^{1} \beta_n^2(x)\ [1 - F_\varepsilon\{\beta_n(x) - \lambda_n\}]\ dx = \int_0^1 X_1(s)\, ds,$$

where $X_1(s) = \int_{1-x_1(n)}^{1-\xi_n(s)} \beta_n^2(x)\, dx$ and $\xi_n(s) = 1 - \beta_n^{-1}\{\lambda_n + Q_{|\varepsilon|}(s)\}$, with $Q_{|\varepsilon|}(s)$ the quantile function of the error's magnitude $|\varepsilon|$. Denote $\zeta_n(s) = \int_0^{1-\xi_n(s)} \beta_n^2(x) dx$, then $X_1(s) = \zeta(s) - x_1(n)$. The function $X_1(s)$ can be verified to be non-decreasing in s and $X_1(1) = 1 - x_1(n)$. For $\widetilde{x}_1(n) \leq 2x_1(n)$, it is thus sufficient that $X_1\{1 - x_1(n)\} \leq x_1(n)$. The analysis of this condition uses the following lemma for $\beta_n(x) \in L_p(r_n)$.

Lemma 7 *For any ξ_n, and $\zeta_n = \displaystyle\int_0^{1-\xi_n} \beta_n^2(x) dx$, we have* $\dfrac{\zeta_n - x_1(n)}{x_1(n) - \xi_n} \leq r_n^{-2p/(2-p)}$.

From the lemma, it follows that $X_1(s) \leq r_n^{-2p/(2-p)}(x_1(n) - \xi_n(s))$. We want, for $s = 1 - x_1(n)$ that $X_1(s) \leq x_1(n)$, which is satisfied if $x_1(n) - \xi_n(s) \leq r_n^{2p/(2-p)} x_1(n)$. We arrive at the condition

$$\frac{\beta_n^{-1}\left[\lambda_n + Q_{|\varepsilon|}\{\beta_n^{-1}(\lambda_n)\}\right] - \beta_n^{-1}(\lambda_n)}{1 - \beta_n^{-1}(\lambda_n)} \leq r_n^{2p/(2-p)}. \tag{33}$$

Condition (33) can be understood as follows: adding the n_1 largest noise component to the n_1 largest signal component does not cause the signal rank order β_n^{-1} to increase substantially.

If Condition (33) is satisfied, then $n_1 = nx_1(n)$ can be taken as model size that meets Assumption 2. As the assumption controls the loss due to missed discoveries, it is automatically satisfied for any larger model $n_1 > nx_1(n)$, while the smallest model $n_1 = nx_1(n)$ tends to $n_1/n \to 0$ thanks to the L_2-concentration in L_p balls.

F Software and reproducible figures

The figures and tables in this paper can be reproduced with routines that are part of the latest version of `ThreshLab`, a Matlab®software package available for download from
`http://homepages.ulb.ac.be/~majansen/software/threshlab.html`.
See

1. `help compareGCVSUREFDRebayesthresh` for Table 1;
2. `help illustrateLARSell0` for Table 2;
3. `help illustratemirroreffect` for Figures 1 and 2;
4. `help comparemirrorpenaltyBirgeMassart` for Figures 3 and 4.