Generalized Cross Validation in variable selection with and without shrinkage

Maarten Jansen
Université Libre de Bruxelles, departments of Mathematics and Computer Science

December 9, 2015

Abstract

This paper investigates two types of results that support the use of generalized cross validation (GCV) for variable selection under the assumption of sparsity. The first type of result is based on the well established links between GCV on one hand and Mallows's C_p and Stein Unbiased Risk Estimator (SURE) on the other hand. The result states that GCV performs as well as C_p or SURE in a regularized or penalized least squares problem as an estimator of the prediction error for the penalty in the neighborhood of its optimal value. This result can be seen as a refinement of an earlier result in GCV for soft thresholding of wavelet coefficients. The second novel result concentrates on the behavior of GCV for penalties near zero. Good behavior near zero is of crucial importance to ensure successful minimization of GCV as a function of the regularization parameter. Understanding the behavior near zero is important in the extension of GCV from ℓ_1 towards ℓ_0 regularized least squares, i.e., for variable selection without shrinkage, or hard thresholding. Several possible implementations of GCV are compared with each other and with SURE and C_p. These simulations illustrate the importance of the fact that GCV has an implicit and robust estimator of the observational variance.

Keywords

Generalized Cross Validation; variable selection; threshold; lasso; Mallows's Cp; sparsity; high-dimensional data;

Correspondence address

Maarten Jansen
Departments of Mathematics and Computer Science,
Université Libre de Bruxelles
Boulevard du Triomphe
Campus Plaine, CP213
B-1050 Brussels - Belgium
maarten.jansen@ulb.ac.be

1 Introduction

The theme of this paper is the application of Generalized Cross Validation (GCV) in the context of sparse variable selection. GCV is an estimator of the predictive quality of a model. Optimization of GCV can thus be used as a criterion to optimize the number of selected variables with respect to the predicting the observations. The size of the selected model can be seen as a smoothing parameter that balances closeness of fit and complexity. Closeness of fit is measured by the residual sum of squares (denoted as SS_E). The complexity of the model, measured by the number of selected variables or an ℓ_p norm of the estimators under the selected model, can be understood as a penalty.

Although GCV has been proposed in quite some situations of sparsity [16, 13] and although it has been analyzed for sparse data [9], the method still needs further theoretical and practical investigation [1]. More specifically, this paper demonstrates that the success of GCV in selecting from sparse vectors rests, not just on some asymptotic optimality, but actually on the combination of two asymptotic results. One result, stated in Proposition 1, focusses on the behavior of GCV in the neighborhood of the optimal smoothing parameter. This value of interest minimizes the risk or expected loss, i.e., the expected sum of squared errors. The result in Proposition 1 then states that, if the data are sufficiently sparse and if this optimal smoothing parameter performs asymptotically well in identifying the significant variables, then the GCV score near the optimal smoothing parameter comes close to the score of Mallows's C_p [11] or Stein's Unbiased Risk Estimator (SURE) [12, 5]. The result is an extension and generalization of previous analyzes [9].

Unlike Mallows's C_p or SURE, GCV does not assume knowledge of the variance of the observational errors. Even in the simple signal-plus-noise model, the implicit variance estimation in GCV is clearly superior to a variance dependent criterion such as C_p or SURE, equipped with a robust explicit variance estimator, as illustrated in Section 5.3.

A second result, stated in Proposition 2, is necessary to ensure that GCV has no global minimum for the full model, i.e., the zero penalty model including all variables. Acknowledgement of the importance of the behavior of GCV near zero is crucial in the extension of GCV beyond its current domains of application. These applications are mostly limited to linear methods [15], typically defined by as an ℓ_2 regularized least squares regression problem, and to ℓ_1 regularized least squares regression problems, i.e., the lasso (least absolute shrinkage and selection operator) [13]. For ℓ_0 regularized problems, the classical definition of GCV cannot be used, because Proposition 2 is not satisfied. A solution for this problem is provided in Section 4, Expression 18.

From the more practical point of view, this paper discusses the remarkable robustness of GCV against violation of the sparsity assumption. We can even use GCV as a variance estimator of an i.i.d. normal vector in the presence of far more than 50% outliers.

This paper is organized as follows. In Section 2 the objectives for the variable selection are stated: the goal is to find a selection that minimizes the prediction error. Definitions are given, together with generalities about unbiased estimators of the prediction error. Section 3 defines GCV, and states an asymptotic result that links GCV to the unbiased estimators of the prediction error. Section 4 states a result about the behavior of GCV for selections close to the full model. The novelty, necessity and consequences of the result are discussed. The working of the resulting variable selection procedures is illustrated in Section 5. Comparison of GCV is made with Stein Unbiased Risk Estimator. The main conclusions are summarized in Section 6. Section 7 contains the proofs of Propositions 1 and 2. Section 8 describes this paper's accompanying software that can be downloaded from the web. The routines allow reproduction of all illustrations used in the text.

2 Unbiased estimators of the prediction error

Consider the classical observational model

$$\boldsymbol{Y} = K \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{y} + \boldsymbol{\varepsilon}, \tag{1}$$

where observations $\boldsymbol{Y}$, noise-free response data $\boldsymbol{y}$ and i.i.d. errors $\boldsymbol{\varepsilon}$ are all n-dimensional real vectors, while the covariates $\boldsymbol{\beta} \in \mathbb{R}^m$, and the design matrix $K \in \mathbb{R}^{n\times m}$. In high-dimensional problems, it is typical to have that $m \gg n$.

The vector $\boldsymbol{\beta}$ is sparse, meaning that most of the variables are zero. The objective is to find and estimate the nonzero values. This presentation of the problem implies that the true model is a subset of the full model.

We investigate estimators that minimize the regularized sum of squared residuals

$$\widehat{\boldsymbol{\beta}}_{\lambda,p} = \arg\min_{\boldsymbol{\beta}} \left[\mathrm{SS_E}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_{\ell_p}^p \right], \tag{2}$$

where the sum of the squared residuals $\boldsymbol{e} = \boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}$ for estimator $\widehat{\boldsymbol{\beta}}$ is defined as $\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}) = \|\boldsymbol{e}\|_{\ell_2}^2$. The regularization in (2) can be interpreted as a constrained optimization problem. This paper restricts discussion to the cases $p = 0$ and $p = 1$, the latter corresponding to the lasso. If $p = 0$, then $\|\boldsymbol{\beta}\|_{\ell_0}^0 = \#\{i \in \{1, \ldots, m\} | \beta_i \neq 0\}$, at least if we set $0^0 = 0$. The estimator is then the minimizer of $\mathrm{SS_E}(\boldsymbol{\beta})$ under the constraint that the number of nonzeros is bounded by a value n_1. The problem of choosing an appropriate value of n_1 is equivalent to choosing the best penalty or smoothing parameter λ. It should be noted that the parameter $\sqrt{\lambda}$ reduces to a hard threshold if $K = I$ (the identity matrix) and $p = 0$ and $\lambda/2$ reduces to a soft threshold if $K = I$ and $p = 1$.

In this paper, the value of the smoothing parameter is optimized with respect to the prediction error. We can write

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p}) = \frac{1}{n} E\|\widehat{\boldsymbol{y}}_\lambda - \boldsymbol{y}\|^2 = \sigma^2 + \frac{E\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda,p})}{n} - \frac{1}{n} 2E(\boldsymbol{\varepsilon}^T \boldsymbol{e}_\lambda). \tag{3}$$

Following [17, 18] we define the degrees of freedom for $\widehat{\boldsymbol{y}}_\lambda$ as

$$\nu_{\lambda,p} = \frac{1}{\sigma^2} E\Big[\boldsymbol{\varepsilon}^T(\boldsymbol{\varepsilon} - \boldsymbol{e}_\lambda)\Big] = n - \frac{E(\boldsymbol{\varepsilon}^T \boldsymbol{e}_\lambda)}{\sigma^2}, \tag{4}$$

then we have

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p}) = \frac{E\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,p})}{n} + \frac{2\nu_{\lambda,p}}{n}\sigma^2 - \sigma^2. \tag{5}$$

This expression is the basis for estimating the prediction error. The term $E\mathrm{SS}_\mathrm{E}$ can be estimated in a straightforward way by SS_E. The variance σ^2 is assumed to be known or easy to estimate. The estimation of the degrees of freedom depends on the model and on the class of estimators under consideration. In all cases and further on in the article, $\boldsymbol{x}$ denotes a binary vector of length m where a one stands for a selected variable and a zero for a non-selected variable. Denote by $K_{\boldsymbol{x}}$ the submatrix of K containing all columns of K corresponding to the selected variables in $\boldsymbol{x}$. Similarly, $\boldsymbol{\beta}_{\boldsymbol{x}}$ stands for the nonzero elements in $\boldsymbol{\beta}$.

- Consider the least squares projection estimator on a submodel $K_{\boldsymbol{x}}$, where the selected set $\boldsymbol{x}$ is independent from the observations. It is well known that in this case the degrees of freedom are $\nu_{\boldsymbol{x}} = n_1$, where n_1 is the number of nonzeros in the selection $\boldsymbol{x}$, at least if $K_{\boldsymbol{x}}$ has full rank. (Otherwise, $\nu_{\boldsymbol{x}} = \mathrm{rank}(K_{\boldsymbol{x}})$.) As the selection $\boldsymbol{x}$ is not driven by p and λ, the degrees of freedom are not indexed by these parameters. Nevertheless $\nu_{\boldsymbol{x}} = n_1$ can be substituted into (5). Then, omission of the expected values, followed by a normalization or Studentization, leads to the classical expression of Mallows's $C_p(n_1) = \mathrm{SS}_\mathrm{E}/n\widehat{\sigma}^2 + 2n_1/n - 1$. If the projection $P_{\boldsymbol{x}}$ onto $K_{\boldsymbol{x}}$ is nonorthogonal, then we find $\nu_{\boldsymbol{x}} = \mathrm{Tr}(P_{\boldsymbol{x}})$.

- In a penalized regression problem, the number of nonzeros $N_{1,\lambda}$ depends on the regularization parameter λ and on the observations. For the lasso, i.e., $p = 1$, it can be proven that $\nu_{\lambda,1} = E(N_{1,\lambda})$. The result holds for any design matrix K, i.e., for both low dimensional [18] and high dimensional [14] data. The proofs motivate its use in the Least Angle Regression (LARS) algorithm [6] for solving the lasso problem. In the signal-plus-noise model, where $K = I$, lasso reduces to soft thresholding. Within this framework, the elaboration of (5) with $\widehat{\nu}_{\lambda,1} = N_{1,\lambda}$ is known as Stein's Unbiased Risk Estimator (SURE) [5].

- In the ℓ_1 regularized case with normal errors, the degrees of freedom depend on $\boldsymbol{\beta}$ only implicitly, through $E(N_{1,\lambda})$, which is easy to estimate. In the ℓ_0 case, even for normal errors, the dependence of $\nu_{\lambda,0}$ on $\boldsymbol{\beta}$ would be explicit, and therefore much harder to estimate. In the case where $K = I$, a quasi unbiased estimation is [7]:

  $$\widehat{\nu}_{\lambda,0} = N_{1,\lambda} + n\int_{-\lambda}^{\lambda}\left(1 - \frac{u^2}{\sigma^2}\right) f_\varepsilon(u)du. \tag{6}$$

  Extensions towards general K are possible [7].

3 The efficiency of Generalized Cross Validation

While GCV can be derived from "classical" (ordinary) leave-out-one cross validation, the analysis in this paper is based on its link with the Mallows's C_p estimation of the prediction error. For any value of p, the non-standardized C_p estimator takes the form

$$\Delta_{\lambda,p} = \frac{1}{n}\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda,p}) + \frac{2\nu_{\lambda,p}}{n}\sigma^2 - \sigma^2. \tag{7}$$

For GCV, this paper uses the following definition

$$\mathrm{GCV}_p(\lambda) = \frac{\frac{1}{n}\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda,p})}{\left(1-\frac{\nu_{\lambda,p}}{n}\right)^2}, \tag{8}$$

which, in practical use, is evaluated by plugging in one of the estimators $\widehat{\nu}_{\lambda,p}$ proposed in Section 2 for $\nu_{\lambda,p}$. The effect of this substitution is limited, as discussed in Section 7.4. The link between GCV and unbiased estimators of the prediction error follows directly from the definitions

$$\mathrm{GCV}_p(\lambda) - \sigma^2 = \frac{\Delta_{\lambda,p} - \left(\frac{\nu_{\lambda,p}}{n}\right)^2\sigma^2}{\left(1-\frac{\nu_{\lambda,p}}{n}\right)^2}. \tag{9}$$

Slightly further developed, this becomes

$$\frac{\mathrm{GCV}_p(\lambda) - \sigma^2 - \Delta_{\lambda,p}}{\Delta_{\lambda,p}} = \frac{\frac{2\nu_{\lambda,p}}{n} - \left(\frac{\nu_{\lambda,p}}{n}\right)^2}{\left(1-\frac{\nu_{\lambda,p}}{n}\right)^2} - \frac{\left(\frac{\nu_{\lambda,p}}{n}\right)^2}{\left(1-\frac{\nu_{\lambda,p}}{n}\right)^2}\cdot\frac{\sigma^2}{\Delta_{\lambda,p}}. \tag{10}$$

For this expression, the following convergence result can be established.

Proposition 1 *Let $\boldsymbol{Y}$ observations of the model in (1), where the number of observations $n \to \infty$. Suppose that the survival function of the errors is bounded by $1 - F_\varepsilon(u) \le L\cdot\exp(-\gamma u)$ for constants γ and L. Denoting $N_{1,\lambda}$ for the number of variables in the model and $n_{1,\lambda} = E(N_{1,\lambda})$, we assume further that there exists a sequence of non-empty sets Λ_n so that $\sup_{\lambda\in\Lambda_n} n_{1,\lambda}\log^2(m)/n \to 0$ for $n \to \infty$. An almost overlapping assumption is made for the degrees of freedom, $\sup_{\lambda\in\Lambda_n}\nu_{\lambda,p} = o(n)$ as $n\to\infty$. Finally, we assume that the estimator $\widehat{\boldsymbol{\beta}}_{\lambda,p}$ has the design coherence property, specified in Assumption 1.*

Under these assumptions, the relative deviation of GCV from $\Delta_{\lambda,p}$ converges to zero in probability. More precisely, denote

$$Q_\lambda = \frac{|\mathrm{GCV}_p(\lambda) - \sigma^2 - \Delta_{\lambda,p}|}{\Delta_{\lambda,p} + V_n}, \tag{11}$$

where V_n is a random variable, independent from λ, and defined by

$$V_n = \max\left(0, \sup_{\lambda\in\Lambda_n}\left(\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p}) - \Delta_{\lambda,p}\right)\right).$$

Then, for $n \to \infty$, $\sup_{\lambda \in \Lambda_n} Q_\lambda \xrightarrow{\text{P}} 0$.

Proof. See Section 7.2.

Assumption 1 ***(Design coherence property)*** *Consider a sequence of models (1), indexed with the sample size n, i.e., $\boldsymbol{Y}_n = K_n \boldsymbol{\beta}_n + \boldsymbol{\varepsilon}_n$. Let $\boldsymbol{K}_{n,i}$ denote the ith row of K_n. Let $\{\widehat{\boldsymbol{\beta}}_{n,\lambda} = 1, 2, \ldots\}$ be a sequence of estimators, depending on a parameter $\lambda \in \Lambda_n$. Let $\Sigma_{n,\lambda}$ be the covariance matrix of $\widehat{\boldsymbol{\beta}}_{n,\lambda}$ and $D_{n,\lambda}$ the diagonal matrix containing the diagonal elements of $\Sigma_{n,\lambda}$, i.e., the variances of $\widehat{\boldsymbol{\beta}}_{n,\lambda}$. Then the sequence of estimators is called design coherent w.r.t. the sequence of parameter sets Λ_n, if there exists a positive c, independent from n, so that for all $\lambda \in \Lambda_n$,*

$$\boldsymbol{K}_{n,i} \Sigma_{n,\lambda} \boldsymbol{K}_{n,i}^T / \boldsymbol{K}_{n,i} D_{n,\lambda} \boldsymbol{K}_{n,i}^T \geq c. \tag{12}$$

Section 7.1 gives an interpretation of this assumption.

The result in Proposition 1 can be understood as follows. The curve of $\mathrm{GCV}_p(\lambda) - \sigma^2$ is a close approximation of the experimental curve of $\Delta_{\lambda,p}$, where both curves are functions of $\lambda \in \Lambda_n$. For use in the forthcoming Corollary 1, the quality of the approximation is expressed in a relative fashion, so that when $E(\Delta_{\lambda,p}) = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p})$ tends to zero for $n \to \infty$, the approximation error vanishes faster in probability. As explained by Corollary 1, this behavior allows us to use the minimizer of GCV as an efficient estimator of the optimal λ. The argument of Corollary 1 requires, however, that the relative approximation error is defined with a denominator that is a vertical shift of $\Delta_{\lambda,p}$, so that this denominator has the same minimizer as $\Delta_{\lambda,p}$. Therefore, the definition of Q_λ in (11) has $\Delta_{\lambda,p} + V_n$ in the denominator rather than $E(\Delta_{\lambda,p})$. A second reason for using $\Delta_{\lambda,p} + V_n$ in the denominator of (11) lies in the proof of Proposition 1. This proof hinges on the close connection between the experimental curves $\mathrm{GCV}_p(\lambda)$ and $\Delta_{\lambda,p}$ in (10). The value of V_n is the smallest vertical shift so that $\Delta_{\lambda,p} + V_n$ majorizes $E(\Delta_{\lambda,p})$ for all $\lambda \in \Lambda_n$. Its value guarantees that $\Delta_{\lambda,p} + V_n \geq \Delta_{\lambda,p}$ and also that $\Delta_{\lambda,p} + V_n \geq E(\Delta_{\lambda,p})$. Both properties are being used throughout the proof of Proposition 1. With $E(\Delta_{\lambda,p})$ as a lower bound, the denominator is also protected against occasionally near-zero or even negative values of $\Delta_{\lambda,p}$.

Corollary 1 thus states that estimating the minimizer of $\Delta_{\lambda,p}$ by the minimizer of $\mathrm{GCV}_p(\lambda)$ may result in a different value for λ, but both, random, values have asymptotically the same quality in terms of $\Delta_{\lambda,p}$, shifted towards its expected value.

Corollary 1 *Let $\widehat{\lambda}_n^* = \arg\min_{\lambda \in \Lambda_n} \Delta_{\lambda,p}$ in the observational model (1) and $\widehat{\widehat{\lambda}}_n = \arg\min_{\lambda \in \Lambda_n} \mathrm{GCV}_p(\lambda)$, with Λ_n defined in Proposition 1, then it holds that*

$$\frac{\Delta_{\widehat{\widehat{\lambda}}_n,p} + V_n}{\Delta_{\widehat{\lambda}_n^*,p} + V_n} \xrightarrow{\text{P}} 1, \tag{13}$$

with V_n as in Proposition 1.

Proof. From the definition of Q_λ in (11), we have for any $\lambda \in \Lambda_n$,

$$-Q_\lambda(\Delta_{\lambda,p} + V_n) \leq \mathrm{GCV}_p(\lambda) - \sigma^2 - \Delta_{\lambda,p} \leq Q_\lambda(\Delta_{\lambda,p} + V_n),$$

which is of course equivalent to

$$(1 - Q_\lambda)(\Delta_{\lambda,p} + V_n) \leq \mathrm{GCV}_p(\lambda) - \sigma^2 + V_n \leq (1 + Q_\lambda)(\Delta_{\lambda,p} + V_n).$$

Since $\widehat{\widehat{\lambda}}_n$ minimizes $\mathrm{GCV}_p(\lambda)$, we have

$$(1 - Q_{\widehat{\widehat{\lambda}}_n})(\Delta_{\widehat{\widehat{\lambda}}_n,p} + V_n) \leq \mathrm{GCV}_p(\widehat{\widehat{\lambda}}_n) - \sigma^2 + V_n \leq \mathrm{GCV}_p(\widehat{\lambda}^*_n) - \sigma^2 + V_n \leq (1 + Q_{\widehat{\lambda}^*_n})(\Delta_{\widehat{\lambda}^*_n,p} + V_n),$$

which leads to

$$1 \leq \frac{\Delta_{\widehat{\widehat{\lambda}}_n,p} + V_n}{\Delta_{\widehat{\lambda}^*_n,p} + V_n} \leq \frac{1 + Q_{\widehat{\lambda}^*_n}}{1 - Q_{\widehat{\widehat{\lambda}}_n}},$$

where the upper bound summarizes the outer inequalities above, while the lower bound follows from the fact that $\widehat{\lambda}^*_n$ minimizes $\Delta_{\lambda_n,p}$. From the convergence of the sequences $Q_{\widehat{\lambda}^*_n}$ and $Q_{\widehat{\widehat{\lambda}}_n}$ and application of Slutsky's theorems we conclude that $\left(\Delta_{\widehat{\widehat{\lambda}}_n,p} + V_n\right) / \left(\Delta_{\widehat{\lambda}^*_n,p} + V_n\right)$ convergences in distribution to 1, which is equivalent to convergence in probability since the limiting variable is a constant.

Remark 1 *Proposition 1 relates GCV to $\Delta_{\lambda,p}$. This relationship is tight in the sense that it holds for every sample separately, not just for the expected values. This is in contrast to the unbiasedness property that links the expected value of $\Delta_{\lambda,p}$ to the prediction error $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p})$. It can easily be verified that the link between $\Delta_{\lambda,p}$ and $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p})$ cannot be stated without the expected value. On the other hand, the proof of Proposition 1 could easily be weakened to an asymptotic optimality for the expected curves. Such a proof would roughly correspond to the first half of the proof of Proposition 1, and reduce to (still) an extension of earlier results [9]. These earlier results covered the simple sparse signal-plus-noise case. The extension of Proposition 1 is thus twofold: the result is stronger and applies to sparse regression with a general design matrix K.*

Remark 2 *The sequence of sets Λ_n must satisfy $\sup_{\lambda \in \Lambda_n} \nu_{\lambda,p}/n \to 0$ as $n \to \infty$. This excludes sequences of λ that would lead to non-sparse selections, i.e., where $\nu_{\lambda,p}/n$ would not vanish. From (10), it can be understood that without $\nu_{\lambda,p}/n$ tending to zero, the relative error of $\mathrm{GCV}_p(\lambda)$ cannot possibly vanish. If the relative size of the true model compared to the sample size does not tend to zero, it cannot be found with increasing efficiency using GCV. Models with large values of $\nu_{\lambda,p}$ correspond to small values of λ. In typical practical applications, the construction of an appropriate sequence of sets Λ_n poses no problem. For instance, in the Gaussian signal-plus-noise model, it is known that the optimal value for λ lies within a constant from the universal value $\lambda_{n,\mathrm{univ}} = \sqrt{2\log(n)}\sigma$ [8]. The sets Λ_n can thus be chosen to be $\Lambda_n = [\gamma\lambda_{n,\mathrm{univ}}, \infty)$, with $\gamma < 1$, or $\Lambda_n = [\lambda_{n,\mathrm{univ}} - \kappa, \infty)$. Although the prediction* *error of large models selected by small values of λ cannot be estimated by GCV in a consistent way and although large models are not of direct interest in a problem of sparse variable selection, it is important that the* $\mathrm{GCV}_p(\lambda)$ *curve behaves appropriately for these cases. This is explained in Section 4. More interpretation of the assumptions in Proposition 1 follow in Section 7.1.*

4 The behavior of GCV near zero penalties

In spite of the asymptotically good behavior of GCV in the neighborhood of the penalty value minimizing the prediction error, the practical application of GCV may suffer from undesired effects outside that neighborhood. In particular, uncareful definition of GCV may result in an expression that tends to zero for zero penalties, i.e., for the full model. That zero value is then a global minimum. This global minimum is likely to hide the true minimum of the prediction error, or at least may hinder the numerical implementation of what becomes a delicate local minimization routine. Examples are given in Figure 4.

This section first formulates a result for the use of GCV in ℓ_1 regularized least squares variable selection, i.e., lasso, in Proposition 2. The proposition states that for lasso, GCV behaves nicely near zero values of λ. Next, the section investigates the behavior of GCV for small values of λ in the case of best $n_{1,\lambda}$-term least squares estimation, this is ℓ_0 regularized variable selection. The application of GCV as defined in (8) is concluded to be problematic, due to the lack of a result like Proposition 2 for ℓ_0 penalties. This motivates the introduction of an alternative definition of GCV for hard thresholding in (18).

Proposition 2 *Given the observational model in (1), with i.i.d. errors, suppose that* $\mathrm{rank}(K) = n$ *with* $n \leq m$. *Also suppose that the cumulative distribution function of the errors is continuous. Let* $\widehat{\boldsymbol{\beta}}_{\lambda,1}$ *be the minimizer of the penalized sum of squared residuals (2) with* $p = 1$, *and consider GCV as defined in (8), then we have*

$$\lim_{\lambda \to 0} E\left[\mathrm{GCV}_1(\lambda)\right] = c_n > 0. \tag{14}$$

In particular, in the signal-plus-noise model, where K *is the identity matrix, we have, for* $n \to \infty$,

$$c_n \to \frac{1}{4 f_\varepsilon^2(0)}, \tag{15}$$

assuming a boundedly differentiable error density function $f_\varepsilon(u)$ *and assuming that sparsity allows one to write that* $\|\boldsymbol{\beta}\|_1 = o(n)$. *In the case of signal plus normal noise, this becomes*

$$c_n \to \sigma^2 \pi / 2. \tag{16}$$

Proof. See Section 7.3

Remark 3 *The result of Proposition 2 confirms the conclusions in Remark 2. Although a sufficiently large value of c_n may prevent the origin, $\lambda = 0$, from being a local minimum of the $\mathrm{GCV}_1(\lambda)$ curve, the value does not correspond to what could be expected from extrapolation of the result in Proposition 2. In particular, for the signal-plus-noise model, the value of c_n in (15) is found to be*

$$c_n = \lim_{\lambda \to 0} E\left[\mathrm{GCV}_1(\lambda)\right] \neq \lim_{\lambda \to 0} E(\Delta_{\lambda,1}) + \sigma^2 = 2\sigma^2.$$

Therefore, Proposition 1 cannot possibly hold for λ close to zero. This motivates the assumption that $\sup_\lambda \nu_{\lambda,p}/n \to 0$ for the sets Λ_n under consideration. We thus find that $\mathrm{GCV}_p(\lambda)$ works fine in the region of interest Λ_n, while the behavior near $\lambda = 0$ is just good enough not to disturb a successful application of the method.

As can be seen from the proof in Section 7.3, the good behavior of $\mathrm{GCV}_1(\lambda)$ is due to the fact that for $\lambda \to 0$, the numerator and the denominator tend to zero at the same rate. The remainder of this section investigates the numerator and the denominator of $\mathrm{GCV}_0(\lambda)$. First, in the case of $n > m$, the system $K\boldsymbol{\beta} = \boldsymbol{Y}$ has no exact solution. As a consequence, for any p, neither the numerator, nor the denominator of $E[\mathrm{GCV}_p(\lambda)]$ tend to zero. For instance, ℓ_0 penalized SS_E works fine with GCV when $n > m$. In other words, good behavior of GCV for models close to the full model is guaranteed if that full model can still be estimated. This may even be the case for rather complex models, as long as the number of observations is large enough; see [10, p.668], where $m < n - 1$.

For high-dimensional data, with full model complexities far beyond the sample size, results as Proposition 2 are crucial. For instance, when $n \leq m$, GCV as defined in (8), does not work with ℓ_0 penalization. Indeed, restricting discussion in the remainder of this section to the signal-plus-noise model, the numerator becomes

$$\frac{1}{n} E\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,0}) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\lambda}^{\lambda} u^2\, f_{Y_i}(u)\, du \asymp \lambda^3.$$

In the denominator $\nu_{\lambda,0}$ can be approximated by $E\widehat{\nu}_{\lambda,0}$, defined in (6), for which it holds that

$$\begin{aligned} 1 - \frac{E\widehat{\nu}_{\lambda,0}}{n} &= 1 - \frac{EN_{1,\lambda}}{n} - \int_{-\lambda}^{\lambda} \left(1 - \frac{u^2}{\sigma^2}\right) f_\varepsilon(u) du \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\lambda-\beta_i}^{\lambda-\beta_i} f_\varepsilon(u) du - \int_{-\lambda}^{\lambda} f_\varepsilon(u) du + \int_{-\lambda}^{\lambda} \frac{u^2}{\sigma^2} f_\varepsilon(u) du \\ &= \frac{1}{n}\sum_{i\in\mathcal{S}_1} \left[\int_{-\lambda-\beta_i}^{\lambda-\beta_i} f_\varepsilon(u) du - \int_{-\lambda}^{\lambda} f_\varepsilon(u) du + \int_{-\lambda}^{\lambda} \frac{u^2}{\sigma^2} f_\varepsilon(u) du\right] + \frac{n - s_1}{n} \cdot \int_{-\lambda}^{\lambda} \frac{u^2}{\sigma^2} f_\varepsilon(u) du. \end{aligned}$$

In the last line, $\mathcal{S}_1$ stands for the set of indices for which $\beta_i \neq 0$ and $s_1 = \#\mathcal{S}_1$. The first term in the last line is of order $\mathcal{O}(\lambda)$ while the second term is of order $\mathcal{O}(\lambda^3)$. As a consequence, the theoretical behavior of GCV is

$$E[\mathrm{GCV}_0(\lambda)] \asymp \frac{\lambda^3}{(\mathcal{O}(\lambda)+\mathcal{O}(\lambda^3))^2} \asymp \lambda \to 0,$$

but, depending on the number and values of the nonzeros in $\boldsymbol{\beta}$, the second term in the denominator may dominate up to very small values of λ, causing $E[\mathrm{GCV}_0(\lambda)]$ to show a large local maximum right after $\lambda = 0$, see Figure 4(b). The presence of this local maximum near the origin (threshold zero) prevents the zero GCV value in the origin from disturbing the local minimization routine near the optimal threshold value. On the other hand, the maximum itself may have an influence on the position of the subsequent local minimum. Moreover, the presence of the local maximum is uncertain. The instability is further enhanced by the fact that $E\widehat{\nu}_{\lambda,0}$, defined in (6), requires knowledge or estimation of σ^2. Such an estimated error variance could be

$$\widehat{\sigma}_{\mathrm{MAD}} = \mathrm{median}|\boldsymbol{Y}|/\Phi_1^{-1}(3/4) \approx \mathrm{median}|\boldsymbol{Y}|/0.6745. \tag{17}$$

The variance estimation may further affect the local maximum and minimum of the GCV curve.

In order to stabilize the estimation, and still working within a signal-plus-noise model, we can use the identity $(1/n)\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) = (1/n)\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,0}) + \lambda^2 N_{1,\lambda}/n$ to rewrite (7) as $\Delta_{\lambda,0} = (1/n)\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) + (2\nu_{\lambda,0}/n)\sigma^2 - \sigma^2 - \lambda^2 N_{1,\lambda}/n$. The first three terms can be approximated in a stable way by a generalized cross validation, leading to an expression for GCV for hard thresholding, i.e., ℓ_0 constrained least squares in a signal-plus-noise model.

$$\mathrm{GCV}_{\mathrm{H}}(\lambda) = \frac{\frac{1}{n}\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,0}) + \lambda^2 \frac{N_{1,\lambda}}{n}}{(1-\frac{\nu_{\lambda,0}}{n})^2} - \lambda^2 \frac{N_{1,\lambda}}{n}, \tag{18}$$

where $\nu_{\lambda,0}$ can be estimated, nearly unbiasedly, as in (6), where the variance σ^2 is estimated by $\widehat{\sigma}^2 = \mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,0})/n$. This pilot estimator is sufficiently accurate near the optimal value of λ.

5 Illustrations and comparative study

This section discusses alternatives for the implementation of GCV and compares the performances of GCV with SURE.

5.1 GCV for solving sparse systems with ℓ_1 penalties

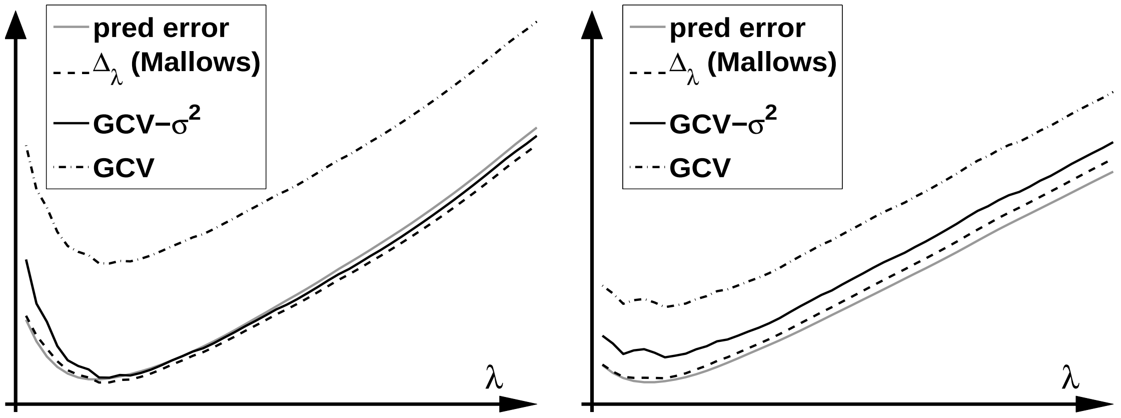
Figure 1(a) presents a typical GCV curve for the outcome of the minimization problem in (2) with $p = 1$ as function of parameter λ. The observations come from the model of (1), where for the simulation, $\boldsymbol{\beta}$ is generated from a zero inflated Laplace (double exponential) distribution, i.e., $P(\beta_i = 0) = 1 - q$ and $\beta_i|\beta_i \neq 0 \sim \mathrm{Laplace}(a)$, where $X \sim \mathrm{Laplace}(a) \Leftrightarrow f_X(x) = (a/2)\exp(-a|x|)$. The observational errors are assumed to be normally distributed. Such a needle-and-haystack model is typical in a Bayesian description of problems involving sparsity. In the simulation, the parameters are set to be $a = 1/5$, $q = 0.01$, $\sigma = 1$. The number of variables equals $m = 3000$, of which approximately $mq = 30$ are nonzero, while the number of observations equals $n = 1000$. The design matrix K is also generated artificially. Obviously, in real problems, many types of K may occur, each with their own characteristics. Covering all sorts of problems would be far beyond the scope of this paper. The simulation in this paper constructs a band limited matrix (with bandwidth equal to 20), with randomly chosen entries within the band. The main motivation for this type of matrix has been the computational feasibility. Solving the lasso problem (2) with $p = 1$ for arbitrary, full matrices is computationally intensive, whether the implementation is based on direct methods such as LARS [6] or on iterative routines such as iterative soft thresholding [3]). The simulation presented here adopts iterative soft thresholding as solver of (2). Figure 1(b) depicts the comparative plots for the same simulation setup, except for the degree of sparsity, which equals $q = 0.05$ instead of $q = 0.01$. Since the vector $\boldsymbol{\beta}$ is less sparse, the values of $\nu_{\lambda,1}$ in the region of interest are larger. As a consequence, the approximation error Q_λ of $\mathrm{GCV}_1(\lambda) - \sigma^2$ w.r.t. $\Delta_{\lambda,1}$ in (10) and (11) is larger. Fortunately, the approximation error is a slowly fluctuating curve as a function of λ. The result is that $\mathrm{GCV}_1(\lambda) - \sigma^2$ does not approximate the value of $\Delta_{\lambda,1}$ (SURE) very well, but the shapes of the two curves are quite similar. In particular, the minimizer of the approximative curve (GCV) has a good quality as an approximation of the minimizer of the prediction error. This allows one to conclude that even in situations that are still far from asymptotics, GCV can often be used quite successfully. For both cases, Figures 1(a) and (b), it can be noted that the plots were made on the interval $\lambda \in [0, \sqrt{2\log(n)}\sigma]$. The right side of the interval corresponds to the universal threshold, a value which is often adopted in conservative sparse variable selection. The plots illustrate the significant gain in prediction error that can be made by data adaptive selection, using either GCV or SURE. Moreover, the universal threshold equally requires a reliable value for σ.

When the sample size is small, say a few tens or hundreds, some caution is needed in reading the GCV curve. Indeed for small sample sizes, the fluctuations near the origin extend into the region of interest. Simple minimization of GCV is not the best strategy. This is reflected in the oscillations in the GCV efficiency curve for sample sizes up to approximately one thousand in Figure 2(b), which was realized with simple minimization, and in the context of a signal-plus-noise model. The conclusions about the required sample sizes would be the same for most general sparse regression problems, with other matrices K, that is. This is illustrated by the experiment with sample size $n = 1000$ in Figure 1 and Table 1. The development of more sophisticated interpretations of the GCV curve, taking into account the presence of fluctuations, is a subject of further research.

The model behind Figure 1 has been simulated 200 times, each time with freshly generated K, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. The quality of estimators is measured by the efficiency. The efficiency of a choice λ_o (o being GCV or any other method) is defined as

$$\mathrm{Eff}(\lambda_o) = \frac{\min_\lambda \mathrm{PE}(\widehat{\boldsymbol{\beta}})}{\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda_o})}. \tag{19}$$

The following table presents the empirical quantiles of the 200 observed efficiencies, for both GCV and



(a) (b)

Figure 1: GCV and SURE curves for $n = 1000$ observations in a model $\boldsymbol{Y} = K\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with K a band limited matrix and $\boldsymbol{\beta}$ a vector of $m = 3000$ variables, of which a small fraction are nonzero. In (a) the fraction of nonzeros is approximately 1%, in (b) the fraction of nonzeros is 5%.

$m = 3000, n = 1000$			
	5%	50%	95%
GCV	0.8517	0.9719	0.9997
SURE	0.9171	0.9837	0.9998

(a)

$m = 300, n = 100$			
	5%	50%	95%
GCV	0.7390	0.9508	0.9996
SURE	0.8403	0.9757	0.9997

(b)

Table 1: (a) Quantiles of observed efficiencies of GCV and SURE in the model behind Figure 1(b). The SURE method was equipped with the exact observational standard deviation. The study is based on 200 simulation runs. (b) Quantiles of observed efficiencies, for the same model as in Figure 1(b), but now with just $m = 300$ variables and $n = 100$ observations in each simulation run.

SURE. The SURE method was equipped with the exact values of the observational standard deviation, which is of course not available in most practical experiments. This advantage explains the difference in efficiency completely, as further explored in Section 5.3, Lemma 1. We also see that most of the loss in efficiency for GCV compared to the ideal SURE takes place in the lower quantiles. That is because sometimes the minimization of GCV gets stuck in false local minima, or that the global minimum GCV is not always the best option, as its location may be shifted by the observational errors. As mentioned before, these effects have less impact when the number of observations increases, when the optimal penalty parameter moves away from zero.

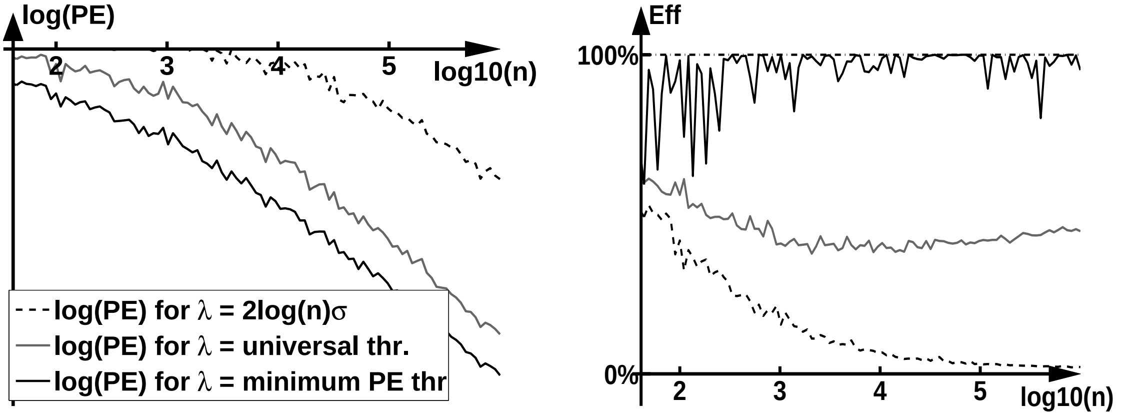
5.2 Asymptotic behavior of GCV

This section investigates the asymptotic behavior of the efficiency defined in (19). For the sake of easy interpretation and discussion, the simulations in this section adopt the signal-plus-noise model, i.e., $K =$

I, although the conclusions can be verified for general K as well. The vector of true parameters $\boldsymbol{\beta}$ is again constructed from a zero inflated Laplacian random variable. As this is an asymptotic study, we let n grow and at the same time, we let proportion of nonzeros decrease as a function of n. This is $P(\beta_i \neq 0) = q(n)$, where we take $q(n) = C\log(n)/n$. We set $C = 10$. Such behavior for $q(n)$ reflects the idea that an increasing number of observations in a nonparametric model allows one to reveal more details. The model becomes richer, which can be seen from the growing absolute number of nonzeros. At the same time, the model becomes sparser, because growing sample sizes lead to an increasing degree of redundancy in the observations. The parameter of the Laplace distribution for the nonzeros also depends on n. In particular, $\beta_i|\beta_i \neq 0 \sim \text{Laplace}\left(a\sqrt{q(n)}\right)$. This way, we have $E(\|\boldsymbol{\beta}\|^2) = n/2a$, while $E(\|\boldsymbol{\varepsilon}\|^2) = n\sigma^2$, keeping the signal-to-noise ratio constant. At the same time, $E(\beta_i^2|\beta_i \neq 0) = 1/2q(n)$, making the nonzeros more prominent against a constant level of measurement noise σ. Any threshold that grows more slowly than the prominence of the nonzeros can be anticipated to asymptotically preserve the signal content of the observations. This is the case for the universal threshold $\lambda_{n,\text{univ}} = \sqrt{2\log(n)}\sigma$. Assuming normal noise, it is a well known result from extreme value theory that the universal threshold asymptotically removes all the noise, meaning

$$\lim_{n\to\infty} P\left(\bigcap_{i=1}^{n}\{|\varepsilon_i| < \lambda_{n,\text{univ}}\}\right) = 1.$$

As a result, the prediction error as defined in (3) tends to zero for large sample sizes. Indeed, both bias and variance contributions to the prediction error tend to zero. The average variance vanishes because the zero β_i's for $n \to \infty$ almost surely do not survive the threshold, leaving them with $\text{var}(\widehat{\beta}_i) \to 0$. The nonzeros are estimated with probability $P(\widehat{\beta}_i = Y_i) \to 1$. The variance in this category tends to σ^2, but the proportion of the nonzeros tends to zero. In a similar way, the average bias can be found to tend to zero. Even larger thresholds, for instance, $\lambda_{n,2} = 2\log(n)\sigma$ are still small enough not to take away essential information from $\boldsymbol{\beta}$ if $n \to \infty$. The minimum prediction error threshold $\widehat{\lambda}_n^*$ is, however, much smaller, as can be verified empirically. Figure 2(a) confirms that all three thresholds result in prediction errors converging to zero. The black solid line has the fastest convergence, it depicts the behavior of the threshold $\widehat{\lambda}_n^*$ with minimum prediction error. The fluctuations in the graphs are due to the fact that for each value of n, we simulate data from a random model for sparsity. Figure 2(b) compares the efficiency of the GCV threshold (in black solid line) with the efficiencies of the universal threshold (in grey solid line) and the threshold $\lambda_{n,2} = 2\log(n)\sigma$ (in black dashed line). It is clear that, although all prediction errors converge to zero, the latter two are not efficient. The GCV threshold focuses on estimating $\widehat{\lambda}_n^*$, explaining its efficiency. The GCV curve has more fluctuations than the other two. This is explained by the fact that in order to minimize GCV, we have to simulate the noise as well, whereas for a fixed threshold, we can compute the prediction error based on the simulated noise-free values of $\boldsymbol{\beta}$.



(a) (b)

Figure 2: (a) Simulated prediction errors as a function of the sample size n for threshold with minimum prediction error $\widehat{\lambda}_n^*$ in black solid line; for the universal threshold $\lambda_{n,\mathrm{univ}} = \sqrt{2\log(n)}\sigma$ in grey solid line; and for the large threshold $\lambda_{n,2} = 2\log(n)\sigma$ in black dashed line. (b) Efficiency as a function of the sample size n, for GCV minimization in black solid line; for the universal threshold in grey solid line; and for the large threshold in black dashed line.

5.3 GCV as variance estimator

This section constructs an estimator of the variance $\mathrm{var}(\varepsilon_i)$ in a signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the noise is assumed to be homoscedastic. The sparse nonzeros in $\boldsymbol{\beta}$ can be seen as the source of outliers in an otherwise identically distributed vector $\boldsymbol{Y}$. The idea is first to estimate $\boldsymbol{\beta}$, using an ℓ_0 or ℓ_1 regularized least squares (2). In the case of a signal-plus-noise model, i.e., $K = I$, the regularization parameter λ becomes a threshold, as mentioned in Section 2. It can be tuned by minimizing $\mathrm{GCV}_p(\lambda)$. The expression of $\mathrm{GCV}_p(\lambda)$ contains a factor $\nu_{\lambda,p}$. If this factor can be estimated without knowing or estimating the variance of the observational errors σ^2, then the evaluation of $\mathrm{GCV}_p(\lambda)$ does not depend on the variance. In that case, the residuals of a GCV driven estimator $\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{GCV}}}$ can be used for the construction of a variance estimator

$$\widehat{\sigma}^2_{\mathrm{GCV}} = \frac{1}{n - \widehat{\nu}_{\lambda_{\mathrm{GCV}}}} \mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda_{\mathrm{GCV}}}) = \frac{1}{n - \widehat{\nu}_{\lambda_{\mathrm{GCV}}}} \sum_{i=1}^{n} (\widehat{\beta}_{i\mathrm{GCV}} - Y_i)^2. \tag{20}$$

In this expression, the subscript GCV refers to the minimum GCV penalty. This adaptively trimmed estimator is remarkably robust against violation of the sparsity assumption. As an example, we compare estimator (20) in a simple threshold scheme for signal-plus-noise observations with the median absolute deviation (MAD) based estimator (17).

The simulation is set up as follows. Given a vector $\boldsymbol{\beta}$ of $n = 100$ unobserved mean values, a proportion of which is zero. The nonzeros have values $\pm M$ with $M = 10\sigma$ and the sign is random. To this vector we add a vector $\boldsymbol{\varepsilon} \sim \mathrm{NID}(\mathbf{0}, \sigma^2)$. (i.e., with normally distributed, independent random

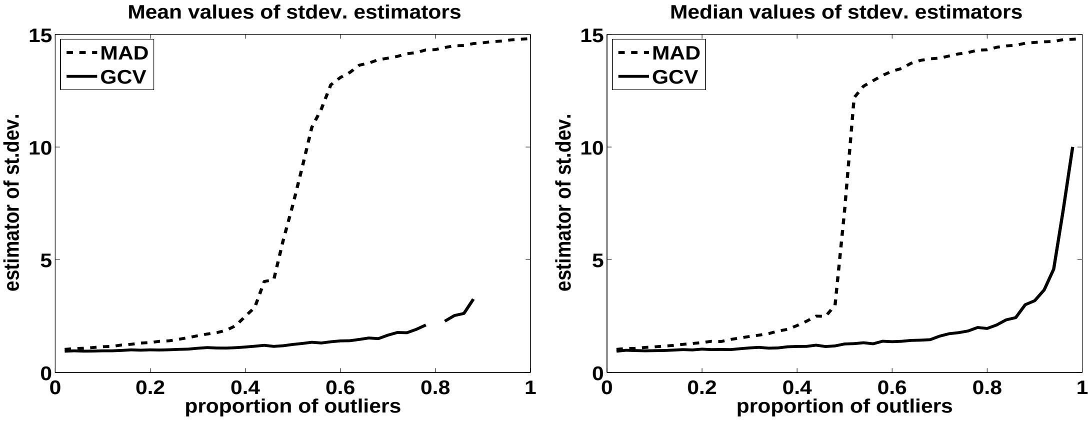


Figure 3: Mean (Left) and median (Right) of 200 estimators $\widehat{\sigma}_{\text{MAD}}$ (Dashed) and $\widehat{\sigma}_{\text{GCV}}$ (Solid line) as a function of the proportion of outliers in otherwise standard normal random variables. The true value of the standard deviation is $\sigma = 1$, which is, of course, almost perfectly estimated when the proportion of outliers is near zero.

variables, all with zero mean and variance σ^2.) From the observed vector $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the variance of $\boldsymbol{\varepsilon}$ is estimated, using $\widehat{\sigma}_{\text{MAD}}$ and $\widehat{\sigma}_{\text{GCV}}$. The GCV based procedure uses soft-thresholding to find $\widehat{y}_{i\text{GCV}}$. The procedure is repeated 200 times for each chosen value of the proportion of nonzeros. The resulting mean and median values of $\widehat{\sigma}_{\text{MAD}}$ and $\widehat{\sigma}_{\text{GCV}}$ are plotted against the proportion of nonzeros in Figure 3.

If the proportion of outliers is larger than 50%, then the GCV threshold trims away more than 50% of the data as well, i.e., $N_{1,\lambda}/n > 50\%$, and moreover, in most cases, the remaining less-than-half of the variables lead to an accurate estimate of the variance.

The GCV based variance estimator also has an interesting property that reveals a sort of feed-back from GCV to SURE/Mallow's C_p. It is stated in the following lemma.

Lemma 1 *Suppose that $\widehat{\boldsymbol{\beta}}_{\lambda,p}$ is continuous w.r.t. λ and let λ_{GCV} minimize $\text{GCV}_p(\lambda)$ as defined in (8). Define a theoretical version of (20), namely $\widehat{\sigma}_o^2 = \text{SS}_\text{E}(\lambda_{\text{GCV}})/(n - \nu_{\lambda_{\text{GCV}},p})$. Then, λ_{GCV} also minimizes $\Delta_{\lambda,p}$, as defined in (7), where σ^2 in that definition is taken to be $\sigma^2 = \widehat{\sigma}_o^2$.*

This feed-back property still holds if the theoretical GCV based variance estimator is replaced by the empirical one, and the definitions of $\text{GCV}_p(\lambda)$ and $\Delta_{\lambda,p}$ adopt the corresponding estimator $\widehat{\nu}_{\lambda,p}$.
Proof. The following lines sketch the proof for the case where $\widehat{\boldsymbol{\beta}}_{\lambda,p}$ is continuously differentiable w.r.t. λ. The case $p = 1$, where the derivative shows discontinuities, can be solved by a continuously differentiable approximation of the ℓ_1-norm. Denote $G(\lambda) = \text{GCV}_p(\lambda) - \widehat{\sigma}^2_{\text{GCV}}$, then if we adopt $\widehat{\sigma}^2_{\text{GCV}}$ in

Definition (7) of $\Delta_{\lambda,p}$, Equation (9) can be rewritten as

$$G(\lambda) = \frac{\Delta_{\lambda,p} - \widehat{\sigma}_o^2 x^2(\lambda)}{[1 - x(\lambda)]^2} \text{ or } \Delta_{\lambda,p} = G(\lambda)\,[1 - x(\lambda)]^2 + \widehat{\sigma}_o^2 x^2(\lambda),$$

with $x(\lambda) = \nu_{\lambda,p}/n$. Taking the derivative leads to

$$\frac{d\Delta_{\lambda,p}}{d\lambda} = G'(\lambda)\,[1 - x(\lambda)]^2 - 2G(\lambda)\,[1 - x(\lambda)]\,x'(\lambda) + 2\widehat{\sigma}_o^2 x(\lambda) x'(\lambda).$$

Substituting $\widehat{\sigma}_o^2 = \mathrm{GCV}_p(\lambda_{\mathrm{GCV}})\,[1 - \nu_{\lambda_{\mathrm{GCV}},p}/n] = G(\lambda)\,[1 - x(\lambda_{\mathrm{GCV}})]\,/x(\lambda_{\mathrm{GCV}})$, yields an expression which is zero if $G'(\lambda) = 0$, thereby completing the proof of Lemma 1. □

The implication of this result is the following. If SURE is equipped with the variance estimator based on GCV, then its efficiency falls back to that of GCV. In other words, the GCV approximation of SURE does not cause any loss in efficiency as such, it only adds an implicit variance estimator, which is of course not perfect, yet surprisingly robust.

5.4 GCV for hard thresholding

This section works within the signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is again an instance from the zero inflated Laplace model, defined in Section 5.1.

$\boldsymbol{\beta}$ is estimated using hard thresholding. A naive implementation for $\mathrm{GCV}_0(\lambda)$ would be to approximate in Expression (8) $\nu_{\lambda,0}$ by $E(N_{1,\lambda})$, the expected number of nonzeros in the selection, which in its turn would be estimated, trivially, by $N_{1,\lambda}$:

$$\mathrm{GCV}_0(\lambda) = \frac{\frac{1}{n}\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,0})}{\left(1 - \frac{E(N_{1,\lambda})}{n}\right)^2}. \tag{21}$$

This corresponds to the dotted lines (0) in Figures 4(a)-(b). The approximation seems to be sufficiently precise in the immediate neighborhood of the optimal threshold as well as for larger threshold values. That is, Proposition (1) probably holds, even if $\nu_{\lambda,0}$ is replaced by $E(N_{1,\lambda})$. Nevertheless, this definition of GCV is useless, since it cannot be used for minimization, due to the global minimum in $\lambda = 0$.

But also a more accurate approximation of $\nu_{\lambda,0}$, using (6) shows practical problems. The corresponding curves carry numbers (1) and (2) in Figures 4(a) and (b). This time, the slopes of a large local maximum disturbs the localization of the true minimum. In the theoretical case where the true value of σ^2 is available, the minimization could still give an acceptable result, see the grey curves (1) in Figures 4(a) and (b). The situation deteriorates when σ^2 has to estimated. Since the variance estimator is sensitive to the degree of sparsity, the problem becomes more prominent when the data are less sparse. Sparsity parameter q is 0.05 in Figure 4(a) and $q = 0.35$ in Figures 4(b) and (c).

The alternative definition in (18), used in Figure 4(c) does not experience nuisance minima and max-

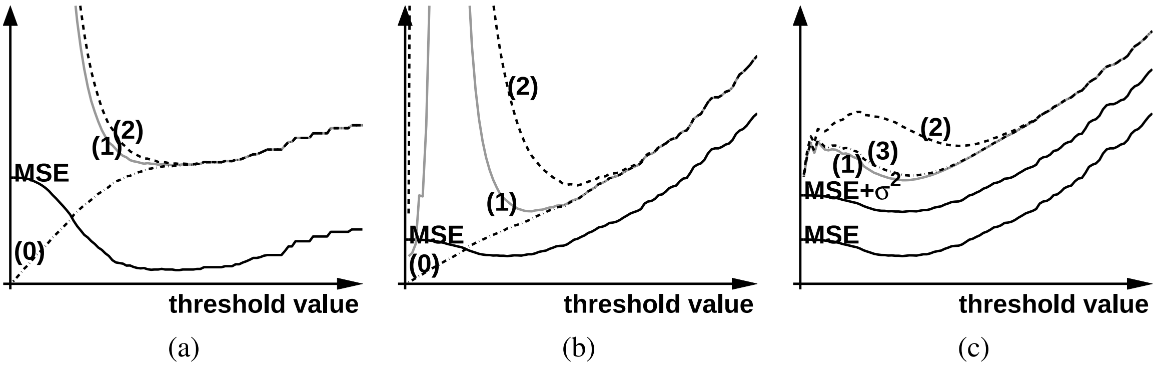


Figure 4: GCV curves for hard-thresholding in signal-plus-noise model. In (a) a fraction of approximately 5% of the noise-free variables are nonzero, in (b) and (c) the fraction is 35%. (a) and (b) use Definition (8) for GCV, except for curves (0), which represent Expression (21). The curves in (c) correspond to Expression (18). Numbers (1), (2), (3) refer to the variance estimation within the GCV definition. (1), i.e., grey lines, stand for exact variance (unknown in practice); (2), i.e., dashed lines, stand for the estimator based on the median absolute deviation (17); (3), i.e., the dotted line in (c), stands for the estimator based on GCV for soft thresholding (20). On all curves, MSE stands for mean squared error, which is $\mathrm{MSE} = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$.

ima. The GCV curve does show a vertical translation w.r.t. what could be expected, i.e., the prediction error (or MSE, mean squared error) plus the variance. This phenomenon is explained in the same way as in Figure 1(b).

6 Summary and concluding remarks

This paper has promoted the use of Generalized Cross Validation (GCV) in the selection of sparse regression models, either with or without shrinkage rules.

Unlike most selection criterions, such as Mallows's C_p or Stein Unbiased Risk Estimator (SURE), GCV does not require an explicit variance estimator. Even in the simple signal plus noise model or in other cases where a variance estimator is easy to construct and robust, the implicit estimation of GCV outperforms the explicit method. In the case of general sparse regression, where the explicit estimation of the variance is more challenging, the benefit of the GCV approach is even more outspoken.

This paper has proven two results in support of GCV. The first result states that GCV mimics the behavior of Mallows's C_p or SURE as a function of the tuning parameter in the neighborhood of the optimal value of that parameter. The approximation by GCV is asymptotically optimal. The contribution of this result compared to previous results is twofold. First, it is found that optimality of GCV holds in probability, and not just for the expected value of GCV. Second, the optimality holds for any sparse regression model, not just for the signal plus noise case (i.e., where the design matrix is the identity matrix).

The second result focusses on the behavior of GCV near zero penalties. This analysis is necessary for successful minimization of GCV. It may even suggest alternative definitions in some applications.

The simulation study has shown that GCV may also be successful in cases that are still far from the asymptotics sketched in the theoretical analysis. Sample size should, however, be large enough. The implicit variance estimator of GCV is more robust than methods where variance is considered as a nuisance parameter. In many applications, the construction of a variance estimator is far from trivial.

7 Proofs of Propositions 1 and 2

7.1 Discussion on the assumptions

The **sparsity** assumptions, stated in Proposition 1, are twofold. First, it is required that $\sup_{\lambda\in\Lambda_n}\nu_{\lambda,p} = o(n)$ for $n\to\infty$, meaning that the number of significant parameters is an order of magnitude smaller than the number of observations. Second, the total number of parameters in the full model, m, must not be too large. Otherwise finding the needles in the haystack becomes too demanding. Similar conditions appear in the literature [2, 4] for successful recovery of sparse data using ℓ_1 regularized least squares methods.

Since the results are presented for a wide class of variable selection and estimation methods, it is not surprising that assumptions are needed on the interaction between the model and its estimation method. The **design coherence property** in Assumption 1, (12) excludes estimators of the model that predict the observations with relatively low variances using highly correlated estimators with arbitrarily large variances. The signal-plus-noise model, where K is the identity matrix, is a prototype of an ideal situation, the constant c in (12) being equal to one. A sufficient but far from necessary condition is that the eigenvalues of the matrix of correlations $D_n^{-1/2}\Sigma_n D_n^{-1/2}$ are bounded from below by c. If the estimator $\widehat{\boldsymbol{\beta}}$ has a diagonally dominated covariance matrix, the assumption is fulfilled. Problems occur when multicollinearity in the covariates is not properly addressed by the regularization. In such a case, a subset of nonzero estimators in $\widehat{\boldsymbol{\beta}}$ can be replaced by a different subset with approximately the same degree of sparsity and nearly the same residuals. A small change in the errors can have a large effect on the selection. The parameter estimation is subject to large volatility, while the effect on the prediction $\widehat{\boldsymbol{y}}$ is limited, because the estimators in $\widehat{\boldsymbol{\beta}}$ have negative correlations: a large value in one subset is compensated by a small or zero value in another subset. As a conclusion, the regularization should compensate for the collinearity in the design.

As a trivial counterexample of design coherence, consider the estimator $\widehat{y}_i = \widehat{\beta}_1 + \widehat{\beta}_2$, so $\boldsymbol{K}_i = [\ 1 \quad 1\ ]$. Then regularization by the restriction that $\widehat{\beta}_1 = -\widehat{\beta}_2$ leads to $\mathrm{corr}(\widehat{\beta}_1, \widehat{\beta}_2) = -1$. This correlation structure is not coherent with the ith row of the design matrix, meaning that the restriction leads to $\mathrm{var}(\widehat{y}_i) = 0$, while $\mathrm{var}(\widehat{\beta}_j)$ for $j \in \{1, 2\}$ can be arbitrarily large.

7.2 Proof of Proposition 1

From Expression (10), and the definition of Q_λ in (11), it is clear that we can concentrate on showing that

$$\sup_{\lambda\in\Lambda_n} \frac{\sigma^2\left(\frac{\nu_{\lambda,p}}{n}\right)^2}{\Delta_{\lambda,p}+V_n} \xrightarrow{\mathrm{P}} 0. \tag{22}$$

Using $E(\boldsymbol{\varepsilon})=\mathbf{0}$, the homoscedasticity of the observational errors, and denoting $\rho_{\lambda,i}=\operatorname{corr}(\varepsilon_i,\varepsilon_i-e_{\lambda,i})$, leads to

$$\begin{aligned}
\nu^2_{\lambda,p} &= \frac{1}{\sigma^4}\left[E\Big(\boldsymbol{\varepsilon}^T(\boldsymbol{\varepsilon}-\boldsymbol{e}_\lambda)\Big)\right]^2 = \frac{1}{\sigma^4}\left[\sum_{i=1}^n \sigma\sqrt{\operatorname{var}(\varepsilon_i-e_{\lambda,i})}\rho_{\lambda,i}\right]^2\\
&\le \frac{1}{\sigma^2}\left[\sum_{i=1}^n \operatorname{var}(\varepsilon_i-e_{\lambda,i})\right]\cdot\left[\sum_{i=1}^n\rho^2_{\lambda,i}\right] \le \frac{1}{\sigma^2}\left[\sum_{i=1}^n E(\varepsilon_i-e_{\lambda,i})^2\right]\cdot\left[\sum_{i=1}^n\rho^2_{\lambda,i}\right]\\
&= \frac{n}{\sigma^2}\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p})\cdot\left[\sum_{i=1}^n\rho^2_{\lambda,i}\right] = \frac{n}{\sigma^2}E(\Delta_{\lambda,p})\cdot\left[\sum_{i=1}^n\rho^2_{\lambda,i}\right].
\end{aligned}$$

The main part of the proof consists in finding an upper bound for $\sup_{\lambda\in\Lambda_n}\sum_{i=1}^n\rho^2_{\lambda,i}$. First, for a given $\lambda\in\Lambda_n$, denote $\rho_{\lambda,ij}=\operatorname{corr}(\varepsilon_i,\widehat{\beta}_{\lambda,j})$, where $\widehat{\beta}_{\lambda,j}$ denotes the jth component of $\widehat{\boldsymbol{\beta}}_{\lambda,p}$. The index p is being omitted as it has no role in the current argument. Then, from $\boldsymbol{\varepsilon}-\boldsymbol{e}_\lambda=\widehat{\boldsymbol{y}}_\lambda-\boldsymbol{y}=K\widehat{\boldsymbol{\beta}}_{\lambda,p}-K\boldsymbol{\beta}$, and using the Cauchy-Schwarz inequality, it follows that

$$\rho_{\lambda,i}=\operatorname{corr}(\varepsilon_i,\widehat{y}_{\lambda,i})=\frac{\sum_{j=1}^m K_{ij}\rho_{\lambda,ij}\sqrt{\operatorname{var}(\widehat{\beta}_{\lambda,j})}}{\sqrt{\operatorname{var}(\widehat{y}_{\lambda,i})}}\le\frac{\sqrt{\sum_{j=1}^m K^2_{ij}\operatorname{var}(\widehat{\beta}_{\lambda,j})}\sqrt{\sum_{j=1}^m\rho^2_{\lambda,ij}}}{\sqrt{\operatorname{var}(\widehat{y}_{\lambda,i})}}. \tag{23}$$

Let $\Sigma_{\widehat{\boldsymbol{\beta}},\lambda}$ denote the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\lambda,p}$ and $D_{\widehat{\boldsymbol{\beta}},\lambda}$ the diagonal matrix whose elements are the diagonal elements of $\Sigma_{\widehat{\boldsymbol{\beta}},\lambda}$. Writing $\boldsymbol{K}_i$ for the ith row of K, the first part of (23) can be rewritten and bounded, for any i and any $\lambda\in\Lambda_n$, as

$$\frac{\sqrt{\sum_{j=1}^m K^2_{ij}\operatorname{var}(\widehat{\beta}_{\lambda,j})}}{\sqrt{\operatorname{var}(\widehat{y}_{\lambda,i})}}=\sqrt{\frac{\boldsymbol{K}_iD_{\widehat{\boldsymbol{\beta}},\lambda}\boldsymbol{K}_i^T}{\boldsymbol{K}_i\Sigma_{\widehat{\boldsymbol{\beta}},\lambda}\boldsymbol{K}_i^T}}\le\frac{1}{c},$$

where the constant $c>0$ has been introduced in (12).

We now establish an upper bound for the remaining factor, $\sqrt{\sum_{j=1}^m\rho^2_{\lambda,ij}}$, in (23). Defining $\xi_{\lambda,j}=\sum_{i=1}^n\rho_{\lambda,ij}\varepsilon_i/\sum_{i=1}^n\rho^2_{\lambda,ij}$, and $\widetilde{\rho}_{\lambda,j}=\operatorname{corr}(\xi_{\lambda,j},\widehat{\beta}_{\lambda,j})$, then $E(\xi_{\lambda,j})=0$, $\operatorname{var}(\xi_{\lambda,j})=\sigma^2$, and $\sum_{i=1}^n\rho^2_{\lambda,ij}=\widetilde{\rho}^2_{\lambda,j}$. Furthermore, denoting $J_{\lambda,j}=\{\widehat{\beta}_{\lambda,j}\neq0\}$, leads to

$$\widetilde{\rho}_{\lambda,j} = \frac{E(\xi_{\lambda,j}\widehat{\beta}_{\lambda,j})}{\sigma\sqrt{\operatorname{var}(\widehat{\beta}_{\lambda,j})}}=\frac{E(\xi_{\lambda,j}\widehat{\beta}_{\lambda,j}|J_{\lambda,j})P(J_{\lambda,j})}{\sigma\sqrt{\operatorname{var}(\widehat{\beta}_{\lambda,j}|J_{\lambda,j})P(J_{\lambda,j})+\left[E(\widehat{\beta}_{\lambda,j}|J_{\lambda,j})\right]^2P(J_{\lambda,j})P(J'_{\lambda,j})}}$$

$$\begin{aligned} &\le \frac{E(\xi_{\lambda,j}\widehat{\beta}_{\lambda,j}|J_{\lambda,j})\,P(J_{\lambda,j})}{\sigma\sqrt{E(\widehat{\beta}^2_{\lambda,j}|J_{\lambda,j})P(J_{\lambda,j})P(J'_{\lambda,j})}} \le \frac{\sqrt{E(\xi^2_{\lambda,j}|J_{\lambda,j})\,E(\widehat{\beta}^2_{\lambda,j}|J_{\lambda,j})}\,P(J_{\lambda,j})}{\sigma\sqrt{E(\widehat{\beta}^2_{\lambda,j}|J_{\lambda,j})P(J_{\lambda,j})P(J'_{\lambda,j})}} \\ &= \frac{\sqrt{E(\xi^2_{\lambda,j}|J_{\lambda,j})}}{\sigma}\sqrt{\frac{P(J_{\lambda,j})}{P(J'_{\lambda,j})}} \end{aligned}$$

For $P(J_{\lambda,j}) \le 1/2$, this becomes $\widetilde{\rho}_{\lambda,j} \le \sqrt{E(\xi^2_{\lambda,j}|J_{\lambda,j})}\sqrt{2P(J_{\lambda,j})}/\sigma$, while for $P(J_{\lambda,j}) \ge 1/2$, it is obvious that $\widetilde{\rho}_{\lambda,j} \le 1 \le \sqrt{2P(J_{\lambda,j})}$. Hence, $\widetilde{\rho}^2_{\lambda,j} \le 2\max(1, E(\xi^2_{\lambda,j}|J_{\lambda,j})/\sigma^2) \cdot P(J_{\lambda,j})$. Define now $p_{\lambda,j} = P(J_{\lambda,j})$, then the event $A_{\lambda,j}$ that maximizes $E(\xi^2_{\lambda,j}|A_{\lambda,j})$ for given $P(A_{\lambda,j}) = p_{\lambda,j}$, is $A_{\lambda,j} = \{|\xi_{\lambda,j}| > Q_{|\xi_{\lambda,j}|}(1 - p_{\lambda,j})\}$, where $Q_{|\xi_{\lambda,j}|}(\alpha)$ is the quantile function of $|\xi_{\lambda,j}|$. Moreover, $E(\xi^2_{\lambda,j}|A_{\lambda,j}) \ge \sigma^2$, so $\max(1, E(\xi^2_{\lambda,j}|J_{\lambda,j})/\sigma^2) \le E(\xi^2_{\lambda,j}|\,|\xi_{\lambda,j}| \ge Q_{|\xi_{\lambda,j}|}(1 - p_{\lambda,j}))/\sigma^2$.

Combination of the upper bounds for the parts of (23) leads to

$$\begin{aligned} \sum_{i=1}^{n} \rho^2_{\lambda,i} &\le \frac{1}{c^2}\sum_{j=1}^{m}\sum_{i=1}^{n}\rho^2_{\lambda,ij} = \frac{1}{c^2}\sum_{j=1}^{m}\widetilde{\rho}^2_{\lambda,j} \le \frac{1}{c^2\sigma^2}\sum_{j=1}^{m} 2\max(\sigma^2, E(\xi^2_{\lambda,j}|J_{\lambda,j}))P(J_{\lambda,j}) \\ &\le \frac{2}{c^2\sigma^2}\sum_{j=1}^{m} E[\xi^2_{\lambda,j}|\,|\xi_{\lambda,j}| \ge Q_{|\xi_{\lambda,j}|}(1 - p_{\lambda,j})]p_{\lambda,j}. \end{aligned} \tag{24}$$

Defining $G_{\lambda,j}(p) = pE(\xi^2_{\lambda,j}|\,|\xi_{\lambda,j}| \ge Q_{|\xi_{\lambda,j}|}(1-p)) = \int_{|x|>Q_{|\xi_{\lambda,j}|}(1-p)} x^2 f_{|\xi_{\lambda,j}|}(x)dx$, it can be verified that $G_{\lambda,j}(p) = \int_0^p Q^2_{|\xi_{\lambda,j}|}(1-\alpha)d\alpha$. The upper bound (24) depends twice on λ, first in the quantile function $Q_{|\xi_{\lambda,j}|}(p)$, and second in its argument. We now take the supremum over the quantile, not yet over its argument. Define $\overline{Q}_{m,n}(p) = \sup_{\lambda\in\Lambda_n}\max_{j=1,\ldots,m} Q_{|\xi_{\lambda,j}|}(p)$ and $\overline{G}_{m,n}(p) = \int_0^p \overline{Q}_{m,n}(1-\alpha)d\alpha$, then $\overline{G}_{m,n}(p)$ is a concave function majorizing the concave functions $G_{\lambda,j}(p)$, while $\overline{G}_{m,n}(0) = 0$. Using the concavity of $\overline{G}_{m,n}(p)$, we thus arrive at

$$\begin{aligned} r(n) &= \sup_{\lambda\in\Lambda_n}\frac{\nu^2_{\lambda,p}}{n^2}\cdot\frac{\sigma^2}{\mathrm{PE}(\boldsymbol{\beta}_{\lambda,p})} \le \frac{1}{n}\sup_{\lambda\in\Lambda_n}\sum_{i=1}^{n}\rho^2_{\lambda,i} \le \frac{2}{c^2\sigma^2}\frac{1}{n}\sup_{\lambda\in\Lambda_n}\sum_{j=1}^{m} G_{\lambda_j}(p_{\lambda_j}) \\ &\le \frac{2}{c^2\sigma^2}\frac{1}{n}\sup_{\lambda\in\Lambda_n}\sum_{j=1}^{m}\overline{G}_{m,n}(p_{\lambda_j}) \le \frac{2}{c^2\sigma^2}\frac{m}{n}\sup_{\lambda\in\Lambda_n}\overline{G}_{m,n}\left(\frac{1}{m}\sum_{j=1}^{m}p_{\lambda,j}\right) \\ &= \frac{2}{c^2\sigma^2}\frac{m}{n}\overline{G}_{m,n}\left(\frac{\sup_{\lambda\in\Lambda_n} n_{1,\lambda}}{m}\right). \end{aligned}$$

At this point, a distinction has to be made according to the behavior of m for $n \to \infty$. If m is constant or weakly depending on n, meaning that $m = \mathcal{O}(\sup_{\lambda\in\Lambda_n} n_{1,\lambda})$ for $n \to \infty$, then $r(n) = \mathcal{O}\{(m/n)\overline{G}_m(\sup_{\lambda\in\Lambda_n} n_{1,\lambda}/m)\} = \mathcal{O}(\sup_{\lambda\in\Lambda_n} n_{1,\lambda}/n)$. For the more common case where m depends strongly on n, Lemma 2 proves that for any m, there exists a value x^*, so that for any $x > x^*$, $[1 - \overline{Q}^{-1}_{m,n}(x)]/L\exp(-\gamma x) \le 1$, where γ and L are constants defined in Proposition 1. Let $p^* = 1 - \overline{Q}^{-1}_{m,n}(x^*)$ and $p = 1 - \overline{Q}^{-1}_{m,n}(x)$. Also let $y = \log(L/p)/\gamma$. Then $L\exp(-\gamma y) = p = [1 - \overline{Q}^{-1}_{m,n}(x)] \le$

$L\exp(-\gamma x)$, and so $y \geq x$, which means $\log(L/p)/\gamma \geq \overline{Q}_{m,n}(1-p)$. All together, for any m, there exists a positive p^* and $0 < p < p^*$, so that $\overline{Q}_{m,n}(1-p) \leq \log(L/p)/\gamma$.

Substituting $p = \sup_{\lambda\in\Lambda_n} n_{1,\lambda}/m \to 0$, and using De l'Hôpital's rule, we find

$$0 \leq \lim_{n\to\infty} r(n) \leq \lim_{n\to\infty} \frac{\int_0^{\sup_{\lambda\in\Lambda_n} n_{1,\lambda}/m}[\log^2(L/p)/\gamma^2]dp}{(\sup_{\lambda\in\Lambda_n} n_{1,\lambda}/m)(n/\sup_{\lambda\in\Lambda_n} n_{1,\lambda})} = \lim_{n\to\infty} \frac{\log^2(Lm/\sup_{\lambda\in\Lambda_n} n_{1,\lambda})/\gamma^2}{n/\sup_{\lambda\in\Lambda_n} n_{1,\lambda}}.$$

The rightmost expression tends to zero if $\sup_{\lambda\in\Lambda_n} n_{1,\lambda}\log^2(m)/n \to 0$.

Finally, we can compute for arbitrary $\delta > 0$,

$$P\left(\sup_{\lambda\in\Lambda_n}\left|\frac{\sigma^2\left(\frac{\nu_{\lambda,p}}{n}\right)^2}{\Delta_{\lambda,p}+V_n}\right| > \delta\right) \leq P\left(\sup_{\lambda\in\Lambda_n}\frac{\sigma^2\left(\frac{\nu_{\lambda,p}}{n}\right)^2}{E(\Delta_{\lambda,p})}\cdot\sup_{\lambda\in\Lambda_n}\frac{E(\Delta_{\lambda,p})}{\Delta_{\lambda,p}+V_n} > \delta\right) \to 0,$$

thereby concluding the proof of Proposition 1. □

Lemma 2 *Let X_i, $i = 1,\ldots,n$ be a collection of independent random variables and suppose that there exists constants γ and L, so that for all $i \in \{1,\ldots,n\} : P(|X_i| \geq x) \leq L\exp(-\gamma x)$. Define*

$$Y_n = \sum_{i=1}^n \alpha_{n,i}X_i,$$

where $\|\boldsymbol{\alpha}_n\|_q \leq 1$, for some $q \in [0,2]$,
then, for any value of n,

$$\lim_{x\to\infty}\frac{P(|Y_n| \geq x)}{e^{-\gamma x}} \leq L. \tag{25}$$

If $\|\boldsymbol{\alpha}_n\|_\infty < 1$, then $P(|Y_n| \geq x) = o\left(e^{-\gamma x}\right)$, as $x \to \infty$, and for any value of n. Let $B_{q,n}$ be a closed ℓ_q unit ball $B_{q,n} = \{\boldsymbol{\alpha}_n | \|\boldsymbol{\alpha}_n\|_q \leq 1\}$ and define $Y_n^* = \sup_{\boldsymbol{\alpha}_n\in B_{q,n}} Y_n$, then Y_n^* satisfies (25).

Proof. Lemma 2 can be proven by induction on n. The case $n = 1$ is trivial. So, suppose that all $\alpha_{n,i}$ are nonzero and that the result (25) holds for $n-1$, then first define $X'_{n-1} = \sum_{i=1}^{n-1}\alpha_{n,i}X_i/\left(\sum_{i=1}^{n-1}|\alpha_{n,i}|^q\right)^{1/q}$. Furthermore, defining $\beta_{n-1} = \left(\sum_{i=1}^{n-1}|\alpha_{n,i}|^q\right)^{1/q} > 0$ and $\beta_n = |\alpha_n| > 0$, $|Y_n|$ can be bounded as $Y_n \leq \beta_{n-1}|X'_{n-1}| + \beta_n|X_n|$. Using the independence of the X_i, it follows that

$$\begin{aligned} P(|Y_n| \geq x) &\leq \int_0^\infty P\left(|X_n| \geq \frac{x-\beta_{n-1}u}{\beta_n}\right)dP(|X'_{n-1}| \leq u) \\ &\leq \int_0^{x/\beta_{n-1}} L\exp\left(-\frac{\gamma(x-\beta_{n-1}u)}{\beta_n}\right)dP(|X'_{n-1}| \leq u) + \int_{x/\beta_{n-1}}^\infty dP(|X'_{n-1}| \leq u) \\ &= L\exp\left(-\gamma x/\beta_n\right)\int_0^{x/\beta_{n-1}}\exp\left(\gamma\beta_{n-1}u/\beta_n\right)dP(|X'_{n-1}| \leq u) + P(|X'_{n-1}| \geq x/\beta_{n-1}). \end{aligned}$$

Since β_{n-1} and β_n are nonzero and positive, and $\beta_{n-1}^q + \beta_n^q = \sum_{i=1}^n |\alpha_{n,i}|^q \leq 1$ we find $\beta_{n-1} < 1$ and

$\beta_n < 1$.

As a result, we have

$$\lim_{n\to\infty} \frac{P(|X'_{n-1}| \geq x/\beta_{n-1})}{e^{-\gamma x}} = \lim_{n\to\infty} \frac{P(|X'_{n-1}| \geq x/\beta_{n-1})}{e^{-\gamma x/\beta_{n-1}}} \frac{e^{-\gamma x/\beta_{n-1}}}{e^{-\gamma x}} \leq L \cdot 0 = 0$$

and, using De l'Hôpital's rule,

$$\begin{aligned} & \lim_{n\to\infty} \frac{L \exp\left(-\gamma x/\beta_n\right) \int_0^{x/\beta_{n-1}} \exp\left(\gamma \beta_{n-1} u/\beta_n\right) dP(|X'_{n-1}| \leq u)}{e^{-\gamma x}} \\ = \; & L \lim_{n\to\infty} \frac{\int_0^{x/\beta_{n-1}} \exp\left(\gamma \beta_{n-1} u/\beta_n\right) dP(|X'_{n-1}| \leq u)}{\exp\left(-\gamma x (1 - 1/\beta_n)\right)} = 0. \end{aligned}$$

We thus conclude that $P(|Y_n| \geq x) = o\left(e^{-\gamma x}\right)$, unless either β_{n-1} or β_n takes the value 1. This situation occurs only if $\boldsymbol{\alpha}_n$ is a Kronecker delta. In that case, the inequality of (25) is trivially satisfied. Uniform convergence over unit ℓ_q-balls can be verified following a similar scheme by induction. □

7.3 Proof of Proposition 2

As $\operatorname{rank}(K) = n$, and $m \geq n$, there exist solutions for the system $K\widehat{\boldsymbol{\beta}}_0 = \boldsymbol{Y}$. Apart from exceptional cases, all solutions have at least n nonzero elements. Selecting the solution $\widehat{\boldsymbol{\beta}}_0$ with smallest value for $\|\widehat{\boldsymbol{\beta}}_0\|_1$ leads to the observation that for $\lambda = 0$ both the numerator and the denominator of $E[\mathrm{GCV}_1(\lambda)]$ are zero.

The numerator of $E[\mathrm{GCV}_1(\lambda)]$ equals $\frac{1}{n} E\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) = \frac{1}{n} E\left[(\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}_{\lambda,1})^T (\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right]$, where $\widehat{\boldsymbol{\beta}}_{\lambda,1}$ is a minimizer of (2) with $p = 1$.

The derivative of the numerator w.r.t. λ is then

$$\begin{aligned} \frac{1}{n} \frac{d}{d\lambda} E\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) &= \frac{1}{n} E\left(\left[\nabla_{\boldsymbol{\beta}} \mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1})\right]^T \cdot \frac{d\widehat{\boldsymbol{\beta}}_{\lambda,1}}{d\lambda} \right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^{n} \left[-2K_i^T (\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right] \frac{d\widehat{\beta}_i}{d\lambda} \right) \end{aligned}$$

The Karush-Kuhn-Tucker conditions for $\widehat{\boldsymbol{\beta}}_{\lambda,1}$ to be the minimizer of (2) impose that $\left(K^T(\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right)_j = \operatorname{sign}(\widehat{\beta}_j) \cdot \lambda$, when $\widehat{\beta}_j \neq 0$ and $\left|\left(K^T(\boldsymbol{Y} - K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right)_j\right| < \lambda$ otherwise. Denoting J_λ the observation dependent index set corresponding to the nonzeros in $\widehat{\boldsymbol{\beta}}_{\lambda,1}$, we have

$$\frac{d}{d\lambda} E\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) = (-2\lambda) \cdot E\left[\sum_{j\in J_\lambda} \operatorname{sign}(\widehat{\beta}_j) \cdot \frac{d\widehat{\beta}_j}{d\lambda}\right] = (-2\lambda) \cdot E\left[\sum_{j\in J_\lambda} \frac{d|\widehat{\beta}_j|}{d\lambda}\right].$$

As $\frac{d}{d\lambda} E\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1})/\lambda$ converges to a nonzero, finite constant when $\lambda \to 0$, it follows that $E\mathrm{SS}_{\mathrm{E}}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) \asymp$

λ^2, where $a(\lambda) \asymp b(\lambda)$ means (here) that $0 < \lim_{\lambda\to 0} a(\lambda)/b(\lambda) < \infty$ (implying the existence of the limit).

The denominator of $E[\mathrm{GCV}_1(\lambda)]$ equals $(1-\nu_{\lambda,1}/n)^2$, where

$$1-\frac{\nu_{\lambda,1}}{n} = \frac{1}{n\sigma^2}E\left[\boldsymbol{\varepsilon}^T\boldsymbol{e}_\lambda\right] = \frac{1}{n\sigma^2}E\left[\boldsymbol{\varepsilon}^T(\boldsymbol{Y}-K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right] = \frac{1}{n\sigma^2}E\left[\boldsymbol{\eta}^T K^T(\boldsymbol{Y}-K\widehat{\boldsymbol{\beta}}_{\lambda,1})\right].$$

The last equation follows from the fact that there must exist a vector $\boldsymbol{\eta}$, independent from λ, for which $\boldsymbol{\varepsilon} = K\boldsymbol{\eta}$, because $\mathrm{rank}(K) = n$. Again the Karush-Kuhn-Tucker conditions allow to write that $1-\frac{\nu_{\lambda,1}}{n} \asymp \lambda$, and so numerator and denominator of $E[\mathrm{GCV}_1(\lambda)]$ are both of order λ^2 for small λ.

In the signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\beta}+\boldsymbol{\varepsilon}$, the limit can be further developed. The denominator equals $(1-\nu_{\lambda,1}/n)^2 = \frac{1}{n^2}\left[\sum_{i=1}^n P(|Y_i|<\lambda)\right]^2$. Given the bounded derivatives of the density $f_\varepsilon(u)$, it holds for small λ that $P(|Y_i|<\lambda) \sim 2\lambda f_{Y_i}(0) = 2\lambda f_\varepsilon(-\beta_i)$. Substitution into the definition of $E[\mathrm{GCV}_1(\lambda)]$ leads to

$$\lim_{\lambda\to 0} E[\mathrm{GCV}_1(\lambda)] = \lim_{\lambda\to 0}\frac{\frac{1}{n}E\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,1})}{4\lambda^2\left[\frac{1}{n}\sum_{i=1}^n f_\varepsilon(-\beta_i)\right]^2} = \lim_{\lambda\to 0}\frac{(1/n)\frac{d}{d\lambda}E\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,1})}{8\lambda\left[(1/n)\sum_{i=1}^n f_\varepsilon(-\beta_i)\right]^2}.$$

As in the signal-plus-noise model $\widehat{\boldsymbol{\beta}}_{\lambda,1} = \mathrm{ST}_\lambda(\boldsymbol{Y})$, with $\mathrm{ST}_\lambda(\boldsymbol{x})$ the soft-threshold function, it holds that $\frac{d|\widehat{\beta}_i|}{d\lambda} = -1, \forall i \in I$, and thus $\frac{d}{d\lambda}E\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,1}) = 2\lambda E(N_{1,\lambda})$, where $N_{1,\lambda}$ is the number of nonzeros in $\widehat{\boldsymbol{\beta}}_{\lambda,1}$. Finally, because of the bounded derivative of the error density function and the sparsity assumption, the denominator can be simplified using

$$\left|f_\varepsilon(0) - \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n f_\varepsilon(-\beta_i)\right| \le \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n |f_\varepsilon(0)-f_\varepsilon(-\beta_i)| \le M\cdot\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n|\beta_i| = 0.$$

In this approximation, M is the upper bound on the absolute derivative of the error density. Substitution into the expression for the limit leads to the result stated in the Proposition. The expression for normal observational errors is a straightforward elaboration, thereby concluding the proof of Proposition 2. □

7.4 The effect of estimated degrees of freedom in the definition of GCV

The definition for GCV in (8), used in this paper, contains the unobserved factor $\nu_{\lambda,p}$. The motivation for adopting a definition with an unobserved non-random factor is that it facilitates the theoretical analysis. Section 2 lists a few cases where this factor can be estimated. We now discuss the effect of the estimation on the quality of GCV as an estimator of the prediction error, leading to the conclusion that the effect is limited. Indeed, let $\mathrm{GCV}_{p,\widehat{\nu}}(\lambda)$ be an empirical analogue of GCV, defined by

$$\mathrm{GCV}_{p,\widehat{\nu}}(\lambda) = \frac{\frac{1}{n}\mathrm{SS}_\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\lambda,p})}{\left(1-\frac{\widehat{\nu}_{\lambda,p}}{n}\right)^2}, \tag{26}$$

then

$$\mathrm{GCV}_{p,\widehat{\nu}}(\lambda) - \mathrm{GCV}_p(\lambda) = \mathrm{GCV}_p(\lambda)\ \frac{(\widehat{\nu}_{\lambda,p}/n - \nu_{\lambda,p}/n)(2 - \widehat{\nu}_{\lambda,p}/n - \nu_{\lambda,p}/n)}{\left(1 - \frac{\widehat{\nu}_{\lambda,p}}{n}\right)^2}.$$

The offset $\mathrm{GCV}_{p,\widehat{\nu}}(\lambda) - \mathrm{GCV}_p(\lambda)$ has an order of magnitude equal to that of $\widehat{\nu}_{\lambda,p}/n - \nu_{\lambda,p}/n$. The offset should be small, compared to Δ_{λ_p} or to $\mathrm{GCV}_p(\lambda) - \sigma^2$. An argument similar to the proof of Corollary 1, then leads to the conclusion that the offset does not perturbate the minimization of $\mathrm{GCV}_p(\lambda)$. As a full analysis would depend on the model and estimation method, we present here a global sketch of the analysis.

The estimator $\widehat{\nu}_{\lambda,p}$ is typically equal or close to the number of nonzeros $N_{1,\lambda}$. If the selection were completely random, then $N_{1,\lambda}$ would be Poisson distributed. This would be the case for $\boldsymbol{\beta} = \mathbf{0}$. Otherwise, in all practical problems where $\boldsymbol{\beta} \neq \mathbf{0}$, the count $N_{1,\lambda}$ is underdispersed, meaning that $\mathrm{var}(\widehat{\nu}_{\lambda,p}) \leq n_{1,\lambda} \approx \nu_{\lambda,p}$. Assuming a negligible bias in $\widehat{\nu}_{\lambda,p}$, we can write

$$\sqrt{\mathrm{var}(\widehat{\nu}_{\lambda,p}/n - \nu_{\lambda,p}/n)} \leq \sqrt{\nu_{\lambda,p}}/n.$$

On the other hand, $E(\Delta_{\lambda_p}) = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\lambda,p})$ is at least of the order $\mathcal{O}(n_{1,\lambda}/n)$. Indeed, even in the ideal case where $\widehat{\boldsymbol{\beta}}_{\lambda,p}$ contains no bias, the expected $n_{1,\lambda}$ nonzero estimators all carry a variance of order at least $\mathcal{O}(1/n)$. As a conclusion, the error $\left\{E\left[(\widehat{\nu}_{\lambda,p}/n - \nu_{\lambda,p}/n)^2\right]\right\}^{1/2} = \mathcal{O}(\sqrt{\nu_{\lambda,p}}/n)$ os slightly smaller than the prediction error itself. More importantly, the the estimator Δ_{λ_p} of the prediction error is inevitably based on the squared residual, and has therefore a standard error of nearly $\mathcal{O}(1/\sqrt{n})$. As a conclusion, the fluctuations in $\mathrm{SS_E}(\widehat{\boldsymbol{\beta}}_{\lambda,p})$ dominate the fluctuations that arise from substitution of $\nu_{\lambda,p}/n$ by $\widehat{\nu}_{\lambda,p}/n$ in (7) or (8).

8 Software and reproducible figures

All 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 simulateCpGCVpaper2012` for Figure 1,

2. `help robustnessGCVvarest` for Figure 3,

3. `help simulateGCVhardthresh` for Figure 4.

9 Acknowledgement

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

References

[1] A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007.

[2] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

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

[4] D. L. Donoho. 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, 2006.

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

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

[7] M. Jansen. Information criteria for variable selection under sparsity. *Biometrika*, 101(1):37–55, 2014.

[8] M. Jansen and A. Bultheel. Asymptotic behavior of the minimum mean squared error threshold for noisy wavelet coefficients of piecewise smooth signals. *IEEE Transactions on Signal Processing*, 49(6):1113–1118, June 2001.

[9] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding. *Signal Processing*, 56(1):33–44, January 1997.

[10] H. Leeb. Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3):661–690, 2008.

[11] C.L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

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

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

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

[15] G. Wahba. *Spline Models for Observational Data*, chapter 4, pages 45–65. CBMS-NSF Regional Conf. Series in Appl. Math. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[16] N. Weyrich and G. T. Warhola. De-noising using wavelets and cross validation. In S.P. Singh, editor, *Approximation Theory, Wavelets and Applications*, volume 454 of *NATO ASI Series C*, pages 523–532, 1995.

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

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