Information criteria bias correction for group selection

Bastien Marquis and Maarten Jansen
Université libre de Bruxelles, departments of Compter Science and Mathematics

December 2021

Abstract

The main contribution of this paper lies in the extension towards group lasso of a Mallows' Cp-like information criterion used in finetuning the lasso selection in a high-dimensional, sparse regression model. The optimisation of an information criterion paired with an ℓ_1-norm regularisation method of the lasso leads to an overestimation of the model size. This is because the shrinkage following from the ℓ_1 regularisation is too permissive towards false positives, since shrinkage reduces the effects of false positives. The problem does not arise with ℓ_0-norm regularisation but this is a combinatorial problem, which is computationally unfeasible in the high-dimensional setting. The strategy adopted in this paper is to select the non-zero variables with ℓ_1 method and estimate their values with the ℓ_0, meaning that lasso is used for selection, followed by an orthogonal projection, i.e., debiasing after selection. This approach necessitates the information criterion to be adapted, in particular, by including what is called a "mirror correction", leading to smaller models. A second contribution of the paper is situated at the methodological level, more precisely in the development of the corrected information criterion using random hard thresholds as a model for the selection process.

Keywords

High-dimension; Sparsity; Variable selection; Mallows' Cp; Group lasso

1 Introduction

In high-dimensional linear regression, where the number of candidate explanatory variables is much larger than the number of observations, the assumption of sparsity is often adopted in the selection of a subset from the candidates for use in subsequent estimation and inference. Simple selection algorithms, such as alternating forward selection and backward elimination become computationally unfeasible in a high-dimensional setting, while offering no guarantee that the outcome is even near to being globally optimal in whatever sense. More advanced algorithms formulate variable selection as a regularised least squares problem, which is equivalent to a constraint minimisation of the sum of the squared residuals over all possible subsets of candidate explanatory variables. Taking the size of the selected subset as constraint leads

to a best k-term orthogonal projection which is combinatorially complex, and hence computationally unfeasible. Instead of the size of the subset, the lasso [26] takes the absolute sum of the estimated values in the selected subset, leading to a convex quadratic optimisation problem, for which many direct and iterative solvers exist, including least angle regression (LARS) [10], interactive soft thresholding [6], and coordinate descent procedures [13, 15]. Furthermore, lasso is variable selection consistent [35, 28, 23], at least under certain conditions, the most important of which amount to stating that the values in the least false model are large enough to be detected. The least false model in this context is the orthogonal projection of the true model (or data generating process, DGP) onto the full or maximal linear model, i.e., the model including all the candidate covariates. Moreover, under the assumption of sparsity and with the right choice of the absolute sum in the constraint, the convex, fast lasso selection mimics the selection by the combinatorial best k orthogonal projection [8].

The topic of this paper lies one step further, at the finetuning of the selection procedure, choosing an appropriate value of k. The choice of k is assessed by the distance between the selected model and the DGP. This distance is measured, for instance, by the Kullback-Leibler (KL) divergence, which amounts to the expected log-likelihood of the selected model under the DGP. The KL divergence is estimated by Akaike's Information Criterion (AIC) [1]. This paper concentrates on the Prediction Error (PE) as distance measure and the corresponding Mallows' Cp [22] as information criterion.

At the level of finetuning the variable selection, lasso and the best k orthogonal projection are not at all equivalent. This follows from the lasso selection including a shrinkage estimation instead of a least squares projection. Shrinkage reduces the effect of falsely selected covariates on the prediction error, echoing Stein's phenomenon [24] that the variance reduction by shrinkage may exceed the introduced bias. Whereas shrinkage may prove beneficial on a fixed selection, finetuning over the size of the selection induces an unpleasant effect. Indeed, by tempering the effect of false positive selections on the prediction error, any method involving shrinkage becomes more tolerant to the presence of these false positives. As a result, the application of shrinkage pushes the minimum of the prediction error curve towards larger models, leading to many false positives.

Alternatives to lasso with concave regularisation, such as SCAD [11], and MCP [33], reduce the shrinkage bias and the tolerance towards false positives, but not completely. The overestimation of the model size can also be remedied by adding data-adaptive weights in the ℓ_1 regularisation, leading to the adaptive lasso [36]. The definition of the weights requires the availability of a prototype estimator, which should be $\sqrt{n}$ consistent for proper working of the procedure. This is the case, for instance, in low dimensional problems where the ordinary least squares solution can be used as prototype.

This paper takes a different approach, following earlier work [17], where the information criterion itself is redeveloped, anticipating for the numerous false positives that may otherwise occur in a high dimensional setting. Lasso is used as a selection algorithm only. The assessment of the lasso selection in the finetuning replaces the shrinkage estimator by a debiased [19] estimator, obtained by least squares projection onto the selected set of covariates. This decoupling of selection and estimation leads to refined information criteria [17]. It comes as no surprise that without the tempering effect of shrinkage, the compromise between false negatives and false positives, and between bias and variance in the prediction, becomes more

delicate, in the sense that small deviations from the optimum may see already substantial increase in either bias or variance.

This could be one of the reasons for working with a unilateral focus on false positives, such as False Discovery Rate control [3] or Knockoffs [12]. Other arguments for a unilateral approach may be application driven. In general, with the refined information criteria as in [17] taking away most of the false positives, information criteria are an interesting choice in applications where both false positives and false negatives are to be avoided. The choice among information criteria can be driven by further compromises, such the compromise between the efficiency of AIC and the consistency of BIC [4], which turn out to be somehow contradictory [30].

The main contribution of this paper consists in the extension of the proposed refined information criteria for use in structured lasso selection. Structure based variable selection methods include the fused lasso [25], the graphical lasso [14] and the composite absolute penalties [34] including the group lasso [32]. Just as in the unstructured case, the adaptive group lasso [29], group SCAD and group MCP [16] provide alternatives to the group lasso in some applications. Other alternatives, in different setups, include two steps procedures for least squares estimation after model selection [2] and recent bootstrap methods in high dimensional lasso [5]. As a second contribution, this paper proposes a novel methodology in the development of the refined information criteria, for use in finetuning structured or unstructured, lasso or other selection routines. More precisely, refined information criteria are found through the introduction of random thresholds on the covariate values, thus modelling the selection process, as explained further below.

The paper is organised as follows. Section 2, explains how to refine Mallows' Cp, anticipating for the false positives during its optimisation. Although the same methodology applies to other information criteria, such as AIC [17], this paper concentrates on Mallows' Cp for a couple of reasons. First, Mallows' Cp has been adopted as a stopping rule in the LARS procedure. Second, it is closely connected to Stein Unbiased Risk Estimator (SURE) [9], used in sparse wavelet coefficient selection. Third, the redeveloped expression of Mallows' Cp remains fairly simple Finally, Mallows' Cp can be linked to Generalised Cross Validation (GCV) [18], which includes an implicit variance estimation. Our main result, presented in Proposition 3.2 in Section 3, provides a workable approximation of the refined Mallows' Cp criterion in a setting of structured signal-plus-noise models. This approximation is further developed for the special case of grouped variables in Section 4. Section 6 provides some simulations and, in Section 7, we compare the efficiency of Mallows' Cp and its bias-corrected version for image denoising using unstructured and group settings.

2 The mirror effect

2.1 The regression model

Consider the regression model with non-random design $\mathbf{X}$,

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

where $\boldsymbol{Z}$ is a n-vector of standardised, independent and identically distributed errors with $\mathrm{var}(Z_i) = 1$ for $i = 1, \ldots, n$. The nuisance parameter σ is assumed to be known or easy to estimate. In the case of identical design $\mathbf{X} = \mathbf{I}$ (signal-plus-noise model), the estimation of σ is straightforward — provided that because of the sparsity most components of $\boldsymbol{\beta}$ are (near-)zero — from the median absolute deviation or interquartile range. In the general case, an implicit variance estimator may be provided by generalised cross validation [18, Section 5.3], whose expression can be seen as a variant of the Mallows' Cp criterion adopted in this paper. The design matrix $\mathbf{X}$ has size $n \times m$ with n possibly smaller than m (high-dimensional data) but we assume that the unknown number n_1 of non-zeros in $\boldsymbol{\beta}$ is smaller than n. Furthermore, we impose that $\boldsymbol{\beta} \in A^m$ where $A^m \subset \mathbb{R}^m$ may be used to model structure amongst the variables. For instance, the variables can belong to groups, or have the form a of tree, or be part of a graphical model. The objective is to find the value k and the corresponding estimator $\widehat{\boldsymbol{\beta}}_k$ with k non-zeros minimising the expected average squared prediction error $\mathrm{PE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n} E\left(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2\right)$, where the prediction is given by $\widehat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_k$. It should be noted here that n_1 represents the true, unknown, theoretic number of nonzeros in the sparse covariate vector $\boldsymbol{\beta}$, while k represents the user-defined size of the selected model. Finetuning k is the topic of the next section.

2.2 Best k selection and mirror effect

Let $\boldsymbol{S}_k$ be the active set, i.e., the set of integers in $\{1, 2, \ldots, m\}$ corresponding to the k non-zeros in $\widehat{\boldsymbol{\beta}}_k$. The notation $\mathbf{X}_{\boldsymbol{S}_k}$ is used for the $n \times k$ submatrix of $\mathbf{X}$ containing the k columns corresponding to the 1s in $\boldsymbol{S}_k$. For a given k, the selection $\boldsymbol{S}_k$ is provided by a procedure $\mathcal{S}_{\mathbf{X}}(\boldsymbol{Y}; k)$, which can be the unstructured best k selection, or any structured procedure. An example of such a procedure could be an implementation of the lasso regularised least squares problem

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + 2\lambda\|\boldsymbol{\beta}\|_1. \tag{2}$$

In this expression, the regularisation parameter λ is finetuned to the sample dependent value $\widehat{\lambda}_k$ defined by the supremum of all values λ leading to an outcome $\widehat{\boldsymbol{\beta}}_k$ with k non-zeros. We investigate the quality of the least squares projection $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k} = (\mathbf{X}_{\boldsymbol{S}_k}^T \mathbf{X}_{\boldsymbol{S}_k})^{-1} \mathbf{X}_{\boldsymbol{S}_k}^T \boldsymbol{Y}$, assuming that $\mathbf{X}_{\boldsymbol{S}_k}$ is non-singular. Let $\boldsymbol{O}_k$ be the selection (i.e., a k-tuple of integers in $\{1, 2, \ldots, m\}$) found by an oracle knowing $\boldsymbol{\mu}$ without noise, using the same procedure as for $\boldsymbol{S}_k$, i.e. $\boldsymbol{O}_k = \mathcal{S}_{\mathbf{X}}(\boldsymbol{\mu}; k)$. Then the least squares projection $\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k} = (\mathbf{X}_{\boldsymbol{O}_k}^T \mathbf{X}_{\boldsymbol{O}_k})^{-1} \mathbf{X}_{\boldsymbol{O}_k}^T \boldsymbol{Y}$ depends on the observations through $\boldsymbol{Y}$, but not through $\boldsymbol{O}_k$. As the selection $\boldsymbol{O}_k$ does not depend on $\boldsymbol{\varepsilon}$, the prediction error $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k})$ is estimated unbiasedly by the non-studentised version of the Mallows' Cp criterion $\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k})$. For a general selection $\boldsymbol{S}$ (here $\boldsymbol{S} = \boldsymbol{O}_k$), the non-studentised Mallows' Cp criterion is given by

$$\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}) = \frac{1}{n} \left\| \boldsymbol{Y} - \mathbf{X}_{\boldsymbol{S}} \widehat{\boldsymbol{\beta}}_{\boldsymbol{S}} \right\|_2^2 + \frac{2|\boldsymbol{S}|}{n} \sigma^2 - \sigma^2. \tag{3}$$

For nearly any procedure $\mathcal{S}_{\mathbf{X}}(\boldsymbol{Y}; k)$, the selection $\boldsymbol{S}_k$ depends on the observations with noise. It is explained below that the least squares estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k}$ on the selection $\boldsymbol{S}_k$ the expectation of $\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})$ in (3) is no longer $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})$. In particular, when the selection $\boldsymbol{S}_k$ follows from minimisation of the information criterion (3), then, under mild conditions detailed in [17],

the prediction error of the oracular selection lies halfway between the true prediction error and the expected value of the information criterion, i.e.,

$$\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k}) - \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k}) = \mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k}) - E\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k}) + o\left(\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})\right). \tag{4}$$

The oracle thus behaves as a "mirror" between the true prediction error and the apparent prediction error provided by the information criterion.

An intuitive explanation for the bias in $\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})$ is the following. When k is small, the selection $\boldsymbol{O}_k$ and $\boldsymbol{S}_k$ will roughly contain the same highly significant variables. By the time k is larger than the true model size n_1, most of the n_1 non-zero variables will be part of the selection. Roughly $\mathcal{O}(k - n_1)$ covariates are chosen to further minimise the residual sum of squares $\|\boldsymbol{Y} - \mathbf{X}_{\boldsymbol{S}_k}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k}\|_2^2$. The selection thus contains false positives that best fit the noise. While the optimisation is meant to best fit the true data, the false positives tend to be those that are furthest away from their true value, which is (near) zero. Choosing the false positives at random would have introduced less noise. Since the selected false positives present themselves as better than a random selection for fitting the response, the Cp finetuning (3) with the mere model size as penalty tends to include them in the minimum Cp selection. As discussed in the introduction, shrinkage estimation would reduce the impact of the false positives on the prediction, however without taking away the false positives from the selection. This appearance versus reality effect can be related to the fact that the optimisation process over random variables $\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}})$ affects the statistics of the selected variables while, for a selection $\boldsymbol{O}$ from an error free oracle, these statistics are left unchanged. It turns out [17] that the oracular curve $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k})$ can be used as a mirror reflecting $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})$ onto $\Delta_p(\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}_k})$.

In the remainder of this paper, $\boldsymbol{S}$ will be associated with a selection size k so we can drop its superscript. We also define $\widehat{\boldsymbol{\mu}}_k = \mathbf{X}_{\boldsymbol{S}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}}$ and note the Mallows' Cp criterion and the prediction error as $\Delta_p(\widehat{\boldsymbol{\mu}}_k)$ and $\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)$ respectively.

2.3 The mirror and degrees of freedom

The effect of false positives on the optimisation of an information criterion can be formalised through the notion of degrees of freedom. Defining the residual vector $\boldsymbol{e}_k = \boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_k$, the generalised degrees of freedom [31] are given by

$$\nu_k = \frac{1}{\sigma^2} E\left[\boldsymbol{\varepsilon}^T(\boldsymbol{\varepsilon} - \boldsymbol{e}_k)\right] = \frac{1}{\sigma^2} E\left(\boldsymbol{\varepsilon}^T\widehat{\boldsymbol{\mu}}_k\right).$$

Straightforward calculations then show that when σ^2 is known and an expression for ν_k can be developed, then

$$\Lambda_p(\widehat{\boldsymbol{\mu}}_k) = \frac{1}{n}\|\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_k\|_2^2 + \frac{2\nu_k}{n}\sigma^2 - \sigma^2 \tag{5}$$

is an unbiased estimate of $\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)$ for any choice of the selection $\boldsymbol{S}$, random or non-random, the degrees of freedom ν_k absorbing any effect of the selection procedure.

Comparison of (3) and (5) reveals that both expressions are identical if $\nu_k = |\boldsymbol{S}| = k$. If the orthogonal projection $\widehat{\boldsymbol{\mu}}_k = \mathbf{P}_{\boldsymbol{S}}\boldsymbol{Y}$ where $\mathbf{P}_{\boldsymbol{S}} = \mathbf{X}_{\boldsymbol{S}}(\mathbf{X}_{\boldsymbol{S}}^T\mathbf{X}_{\boldsymbol{S}})^{-1}\mathbf{X}_{\boldsymbol{S}}^T$ is adopted in the setting of

a given, fixed model $\boldsymbol{S}$, as an estimator of the response $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, then it is straightforward to verify that indeed $\nu_k = k$. If, however, $\boldsymbol{S}$ depends on the sample, as in the case of a variable selection procedure, then the offset between the unbiased estimator of the prediction error in (5) and the Cp criterion in (3) is given by $2(\nu_k - k)\sigma^2/n$, which is, as seen in (4), approximately twice the offset with respect to the oracular prediction error.

The offset comes from the false positives in the selection, which have been chosen to best fit the errors. Reducing the impact of false positives, the shrinkage in the lasso estimator provides an exact compensation of the offset, both in the low dimensional [37] as in the high dimensional [27] case. Compensating the offset in degrees of freedom or in prediction error, shrinkage does not resolve the false positive selections themselves.

Under sparsity assumptions, detailed in [17], the expression of the degrees of freedom in least squares projection after selection can be approximated by

$$\nu_k = \frac{1}{\sigma^2} E\left[\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{\varepsilon}\|_2^2\right] + o\left[\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right] \text{ as } n \to \infty, \tag{6}$$

which reflects exactly the essence of the mirror effect: the offset $\nu_k - k$ is due to the interaction between the noise $\boldsymbol{\varepsilon}$ and the estimator $\mathbf{P}_{\boldsymbol{S}}$ within the sample dependent selection $\boldsymbol{S}$. The approximation motivates the formal definition of the mirror,

$$m_k = \frac{1}{n} E\left[\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{\varepsilon}\|_2^2 ; \boldsymbol{\beta}\right] - \frac{k}{n}\sigma^2. \tag{7}$$

This expression explicitly writes the parametric dependence on $\boldsymbol{\beta}$. From (6) it follows that $m_k = \frac{1}{n}(\nu_k - k)\sigma^2 + o\left[\mathrm{PE}\left(\widehat{\boldsymbol{\mu}}_k\right)\right]$. Calculation, if at all possible, or otherwise estimation or approximation of m_k leads to a workable information criterion, starting from (5), which becomes

$$\Lambda_p(\widehat{\boldsymbol{\mu}}_k) = \frac{1}{n}\left\|\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_k\right\|_2^2 + \frac{2k}{n}\sigma^2 + 2m_k - \sigma^2, \tag{8}$$

In (8), m_k is the exact correction term, an approximation $\widetilde{m}_k$ or an estimator $\widehat{m}_k$, of which the following sections provide a couple of concrete cases.

3 The mirror effect in structured signal-plus-noise models

This section concentrates on the simple signal-plus-noise model. It will construct in Proposition 3.2 an approximation for the mirror m_k based on observable quantities.

3.1 The model and assumptions

Consider the signal-plus-noise model $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \boldsymbol{\beta} + \sigma\boldsymbol{Z}$, that is Model (1) where $\mathbf{X} = \mathbf{I}$. The least squares estimator $\widehat{\boldsymbol{\mu}}_k$ on the active set $\boldsymbol{S}_k$ is given by $i \in \boldsymbol{S}_k \Rightarrow \widehat{\mu}_{k,i} = Y_i$. We assume

(A1) $\boldsymbol{Z}$ is independent and identically distributed with zero mean.

(A2) The density of the noise σZ is unimodal and symmetric around its (zero) mean. The density f_Z has a bounded derivative and $f_Z(t) = o(t^{-4})$ for $t \to \pm\infty$. This assumption means that the tails of the noise distribution are not too heavy. This is because heavy tails induce outliers mixing up with the true significant peaks in $\boldsymbol{\mu}$.

(A3) **Asymptotic sparsity:** the data should be sparse. The precise formulation of this assumption is postponed until few additional notions have been introduced (see Section 3.3).

For the (structured) selection procedure $\mathcal{S}_{\mathbf{X}}(\boldsymbol{Y}; k)$ with given k, we assume

(A4) the selection of component i depends on the absolute value $|Y_i| = |\mu_i + \varepsilon_i|$ only, not on its sign. Moreover, we assume that if the value $|y_i|$ in a given context $\boldsymbol{Y}$ is large enough to be selected, then any value above $|y_i|$ would also be selected in that context.

3.2 Writing the selection event in terms of a random threshold

The rationale behind Assumption (A4) is the following. In an individual variable selection procedure, a threshold on a coefficient's magnitude decides whether that coefficient is in or out. Although in a group selection, the individual magnitude of a coefficient is no longer enough to decide, it makes sense to impose at least that if a coefficient has been selected, along with other members of its group, it would also be selected if its magnitude were larger, the other members remaining unchanged. Conversely, if a coefficient is not selected, then it would remain inactive if its magnitude were diminished, keeping values of the other members of its group.

As a result of Assumption (A4), the selection of component i can be written as the indicator function of a random set evaluated in the observation Y_i, so that

$$P(i \in \boldsymbol{S}; \boldsymbol{\mu}) = P\left(\mathrm{I}_{[T_{k,i},\infty)}(|Y_i|) = 1\right) = P(|Y_i| \geqslant T_{k,i}),$$

with $T_{k,i}$ a positive random variable. In some routines, the closed boundary at $T_{k,i}$ and the inequality should be replaced by an open boundary and a strict inequality, however without any impact on the subsequent discussion. The random set, $[T_{k,i}, \infty)$, depends on a random threshold $T_{k,i}$. The randomness of $T_{k,i}$ is due to its dependence on the other observations $Y_{i'}$ with $i' \neq i$. This dependence may follow from the grouping in a structured selection, but also from the ordering of the absolute values in simple non-structured approach.

In a similar way, we can write

$$E(\varepsilon_i^2 | i \in \boldsymbol{S}_k; \boldsymbol{\mu}) = E\left(\varepsilon_i^2 | |Y_i| \geqslant T_{k,i}\right).$$

Denote by $\mathrm{I}(x \in \boldsymbol{A})$ the indicator function of the set $\boldsymbol{A}$, meaning that $\mathrm{I}(x \in \boldsymbol{A}) = 1$ if $x \in \boldsymbol{A}$ and $\mathrm{I}(x \in \boldsymbol{A}) = 0$ otherwise. Then, in the signal-plus-noise model, the norm of the projection in (7) is given by $\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{\varepsilon}\|_2^2 = \sum_{i=1}^n \varepsilon_i^2 \mathrm{I}(i \in \boldsymbol{S})$ Conditioning on $T_{k,i}$ leads to

$$m_k \;=\; \frac{1}{n}\sum_{i=1}^{n} E(\varepsilon_i^2 \mathrm{I}(i \in \boldsymbol{S}); \boldsymbol{\mu}) - \frac{k}{n}\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} E\left[(\varepsilon_i^2 - \sigma^2)\mathrm{I}(i \in \boldsymbol{S}); \boldsymbol{\mu}\right]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty E\left[(\varepsilon_i^2-\sigma^2)\mathrm{I}(i\in\boldsymbol{S})\middle|T_{k,i}=t;\boldsymbol{\mu}\right]f_{T_{k,i}}(t)dt\\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty E\left[(\varepsilon_i^2-\sigma^2)\mathrm{I}(i\in\boldsymbol{S})\middle|T_{k,i}=t;\mu_i\right]f_{T_{k,i}}(t)dt.
\end{aligned}
$$

In this expression we used the fact that, given the threshold, the selection of Y_i, i.e., $i \in \boldsymbol{S}$ depends parametrically on the value μ_i, not on the other components in $\boldsymbol{\mu}$. This follows from Assumption (A4), which states that the selection of Y_i depends on its absolute value exceeding a threshold fixed by the other components in $\boldsymbol{Y}$. After observation of the threshold, the values of the other components have no further influence on the selection.

The expression can be further developed by writing

$$
m_k = \frac{1}{n}\sum_{i=1}^{n} E\left[h(T_{k,i};\mu_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty h(t;\mu_i)f_{T_{k,i}}(t)dt,
$$

where

$$
\begin{aligned}
h(t;\mu_i) &= E\left[(\varepsilon_i^2-\sigma^2)\mathrm{I}(i\in\boldsymbol{S})\middle|T_{k,i}=t;\mu_i\right] = \int_{-\infty}^{\infty}(u^2-\sigma^2)\mathrm{I}(|\mu_i+u|\geqslant t)f_\varepsilon(u)du\\
&= \int_{-\infty}^{\infty}(u^2-\sigma^2)\mathrm{I}(u\leqslant -t-\mu_i \text{ or } u\geqslant t-\mu_i)f_\varepsilon(u)du
\end{aligned}
$$

Defining $G_\varepsilon(x) = \int_{-\infty}^{x}(u^2-\sigma^2)f_\varepsilon(u)du$, for which the limits in $\pm\infty$ are zero, and using the symmetry of $f_\varepsilon(u)$ for $G_\varepsilon(-x) = -G_\varepsilon(x)$, we can write

$$
h(t;\mu_i) = G_\varepsilon(\infty) - G_\varepsilon(t-\mu_i) + G_\varepsilon(t-\mu_i) - G_\varepsilon(-\infty) = -G_\varepsilon(t-\mu_i) - G_\varepsilon(t+\mu_i), \tag{9}
$$

and from there

$$
m_k = \frac{1}{n}\sum_{i=1}^{n} E\left[-G_\varepsilon(T_{k,i}-\mu_i) - G_\varepsilon(T_{k,i}+\mu_i)\right]. \tag{10}
$$

This expression still depends on the unknown μ_i. Section 3.4 will develop an approximation of (10). For this approximation, we rely on the assumption of sparsity (A3), detailed in Section 3.3.

3.3 Formal assumption on asymptotic sparsity

Now we are at the point where we can formalise the Assumption (A3) on asymptotic sparsity. For this, we need to consider n to grow to ∞. With growing n, the distributions of the random thresholds $T_{k,i}$ and the values of the components in $\boldsymbol{\mu}$ evolve. In order to avoid overloaded notations, the dependence on n in $T_{k,i}$ and $\boldsymbol{\mu}$ is not made explicitly in the formulation of the assumptions or throughout the proof.

By way of standardisation, let σ be independent from the sample size n. We impose that the signal-to-noise ratio remains approximately constant, meaning that there exist positive constants c and C, not depending on n, so that

$$
c \leqslant \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \leqslant C.
$$

As n grows, the vector $\boldsymbol{\mu}$ is assumed to become sparser, concentrating the information in a small (of $o(n)$, that is) number of significant components, thus allowing the procedure to perform a precise selection under optimal or near-optimal value of k. This means that the non-selected parameters in β do not carry much of the information. In the ideal case, the non-selected values are all zero. Otherwise the omission of these values is supposed not to dominate the prediction error. For the proof of the forthcoming proposition, we will need a slightly stricter version, in the sense that also true values a bit above the threshold should not dominate the prediction error. More precisely, we impose that

$$\frac{1}{n}\sum_{i=1}^{n}\mu_i^2 P\left(|\mu_i| < 2T_{k,i}\right) = o\left[\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right]. \tag{11}$$

This corresponds approximately to Assumption 2 in [17]. At the same time, we assume that the random thresholds tend (slowly) to infinity as n grows larger, more precisely

$$\lim_{n\to\infty}\max_{1\leqslant i\leqslant n} E\left(\frac{1}{T_{k,i}}\right) = 0. \tag{12}$$

Assumptions (12) and (11) are somehow contradictory, in the sense that (12) expresses an increasingly strict condition for a parameter to be selected, while (11) imposes that the not selected parameters represent a decreasingly relative contribution to the squared bias. Together, these conditions are met by data where the information is more and more concentrated in a limited number of large components, when n tends to ∞.

Concrete examples of sparse vectors $\boldsymbol{\mu}$ satisfying the assumptions can be constructed in ℓ_p balls with $p < 2$. A prototype concerns a vector of length n of which a small proportion, typically $n_1 = \mathcal{O}(\log(n))$ is non-zero. The non-zeros are supposed to be normalised in ℓ_2 in the sense that $\|\boldsymbol{\mu}\|_2^2/n = 1$. Assuming additive, normal, independent errors ε, the thresholds in this simple setup can be taken to have the universal (i.e., data-independent) value $T_{k,i} = \sqrt{2\log(n)}\sigma$, for which a classical result in extreme value theory [7, 20] states that asymptotically all false positives are eliminated, more precisely $\lim_{n\to\infty} P\left(\max_{i=1,\ldots,n}|\varepsilon_i| > \sqrt{2\log(n)}\sigma\right) = 0$. The threshold satisfies (12). On the other hand, the threshold grows slower than the non-zero values, whose root mean squared value is of the order $\|\boldsymbol{\mu}\|_2^2/\log(n) = n/log(n)$, thus trivially satisfying (11). More realistic examples beyond the prototype can be constructed by the introduction of an index of sparsity generalising the abrupt transition between n_1 zeros and $n - n_1$ non-zeros. Construct an invertible, non-decreasing, positive function $\mu_n(x)$ on $[0, 1]$ so that the ordered elements of $\boldsymbol{\mu}$ are found as $\mu_n(i/n)$ for $i = 1, \ldots, n$, then the index of sparsity (or concentration) can be defined as the value $x_1(n)$ so that

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

Then, as developed in the supplementary material, Section 5, for [17], the concentration index $x_1(n) \to 0$ as $n \to \infty$ can be analysed to provide a framework in which the conditions can be satisfied. On the other hand, ℓ_p-balls with sufficiently small radii provide the right, small enough, values of the concentration index.

3.4 Definition and asymptotic behaviour of the approximation

At this point, we introduce an approximation for the mirror function.

Definition 3.1

$$\begin{aligned}\widetilde{m}_k &= \frac{1}{n}\sum_{i=1}^{n} E\left[(\varepsilon_i^2-\sigma^2)\mathrm{I}(i\in \boldsymbol{S});\boldsymbol{\mu}^{[-i]}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n} P(i\in \boldsymbol{S};\boldsymbol{\mu}^{[-i]})\left[E(\varepsilon_i^2|i\in \boldsymbol{S};\boldsymbol{\mu}^{[-i]})-\sigma^2\right],\end{aligned}$$

where $\boldsymbol{\mu}^{[-i]}$ is the vector obtained by replacing μ_i in $\boldsymbol{\mu}$ by zero.

Using the notations in (9), this becomes

$$\widetilde{m}_k = \frac{1}{n}\sum_{i=1}^{n} E\left[-2G_\varepsilon(T_{k,i})\right].$$

Then, for this approximation, we have the following result stating that replacing $\boldsymbol{\mu}$ by $\mathbf{0}$ in expression (10) lets $\widetilde{m}_k$ perform asymptotically as well as m_k. This way, we can construct the mirror based on observable quantities, without knowing the distribution of $\boldsymbol{\mu}$.

Proposition 3.2 *(main result) Under the aforementioned assumptions, it holds that for $n\to\infty$*

$$|\widetilde{m}_k - m_k| = o\left[\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right].$$

The proof is given in Appendix A.

Corollary 3.3 *Define the information criterion*

$$\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_k) = \frac{1}{n}\left\|\boldsymbol{Y}-\widehat{\boldsymbol{\mu}}_k\right\|_2^2 + \frac{2k}{n}\sigma^2 + 2\widetilde{m}_k - \sigma^2. \tag{13}$$

Let $\widetilde{k}$ be the model size that minimises $E\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_k)$ and let k^* minimise $E\Lambda_p(\widehat{\boldsymbol{\mu}}_k)=\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)$. Then it holds that

$$\frac{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{\widetilde{k}})}{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{k^*})}\to 1. \tag{14}$$

In other words, replacing m_k by $\widetilde{m}_k$ may lead to different selections, even asymptotically, but these selections attain the same balance between false negatives and false positives in terms of prediction error. **Proof.** Denoting

$$\widetilde{\rho}_k = \frac{|\widetilde{m}_k - m_k|}{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)} = \frac{\left|E\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_k)-\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right|}{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)},$$

we have $(1-\widetilde{\rho}_k)\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\leqslant E\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_k)\leqslant(1+\widetilde{\rho}_k)\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)$, and so

$$(1-\widetilde{\rho}_{\widetilde{k}})\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{\widetilde{k}})\leqslant E\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_{\widetilde{k}})\leqslant E\widetilde{\Lambda}_p(\widehat{\boldsymbol{\mu}}_{k^*})\leqslant(1+\widetilde{\rho}_{k^*})\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{k^*}).$$

Together with the fact that $\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{k*}) \leqslant \mathrm{PE}(\widehat{\boldsymbol{\mu}}_{\widetilde{k}})$ this leads to

$$1 \leqslant \frac{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{\widetilde{k}})}{\mathrm{PE}(\widehat{\boldsymbol{\mu}}_{k*})} \leqslant \frac{1+\widetilde{\rho}_{k*}}{1-\widetilde{\rho}_{\widetilde{k}}}.$$

By Proposition 3.2, $\widetilde{\rho}_k = o(1)$, thus completing the proof. □

Provided that the values of the thresholds $T_{k,i}$ can be computed from the sample responses, the approximative mirror is estimated unbiasedly by its empirical counterpart,

$$\widehat{m}_k = \frac{-2}{n} \sum_{i=1}^{n} G_\varepsilon(T_{k,i}), \tag{15}$$

leading to an estimator of the degrees of freedom $\widehat{\nu}_k = k + n\frac{\widehat{m}_k}{\sigma^2}$ and a modified Mallows' Cp criterion as in (13). The approach with the random thresholds will be further developed in Sections 4 and 5 for group lasso selection.

3.5 Estimating the variance

The information criterion (13) assumes knowledge or selection-independent estimation of the variance. In the sparse signal-plus-noise model, such an estimation is provided by the median absolute deviation,

$$\mathrm{MAD}(\boldsymbol{Y}) = \mathrm{median}\left[|Y_i - \mathrm{median}(\boldsymbol{Y})|\right].$$

Based on the statistics of the error vector ε, one can construct a variance estimator. In the case of normal errors, for instance, a well known estimator is given by

$$\widehat{\sigma} = \mathrm{MAD}(\boldsymbol{Y}) \cdot \Phi^{-1}(3/4) = 0.6745 \cdot \mathrm{MAD}(\boldsymbol{Y}),$$

where $\Phi(x)$ denotes the standard normal CDF. In the linear model beyond the signal-plus-noise case, discussed in further detail in Section 5, a possible strategy to estimate σ^2 is to work with a pilot estimator of $\boldsymbol{\beta}$ using the lasso, including the shrinkage, and with generalised cross validation [18] as a information criterion instead of Mallows' Cp. This approach does not rely on an explicit variance estimation. Although the resulting pilot estimator has many false positives, its residual vector may serve in a variance estimator.

4 Correction in signal-plus-noise models with group structure

In this section, we develop the approximation $\widetilde{m}_k$ for the case of group-structured selection.

4.1 Group selection

Let $\boldsymbol{\mu}$ be a sparse vector of groups of variables. The key property for group selection is that all variables from a same group should become non-zeros (or zeros) simultaneously. Considering group lasso as our selection procedure, it is worth noting that the largest groups are more likely

to be included in the model compared to groups of smaller sizes. In their original proposal [32], Yuan and Lin recommended penalising the groups according to their size. In order to simplify calculations, we consider only groups of the same size w. In that case, each group has the same probability of being selected. The number of variables is then equal to $n = r \times w$, with r the number of groups and we can write $\boldsymbol{\mu}$ as the vector of groups $(\boldsymbol{\mu}_j)_{j=1,\dots,r}$. Consequently, we have $(\boldsymbol{Y}_j)_{j=1,\dots,r} = (\boldsymbol{\mu}_j)_{j=1,\dots,r} + (\boldsymbol{\varepsilon}_j)_{j=1,\dots,r}$.

To stay consistent with the previous notations, we define l as the number of selected groups so that $l \times w = k$ is the number of variables included in the model, corresponding to the selection $\boldsymbol{S}$. We find the least squares estimator for a group of variables from $\boldsymbol{S}_j = \mathbf{1}_j \Leftrightarrow \widehat{\boldsymbol{\mu}}_{k,j} = \boldsymbol{Y}_j$, where $\boldsymbol{S}_j$ and $\widehat{\boldsymbol{\mu}}_{k,j}$ are the jth groups from $\boldsymbol{S}$ and $\widehat{\boldsymbol{\mu}}_k$ respectively. The best l group selection, measured by the Cp-value, consists of the l groups from $\boldsymbol{Y}$ whose ℓ_2-norms are the largest. This means that the threshold selecting l groups in a group lasso procedure is the $(r-l)$th order statistic of the group norms of $\boldsymbol{Y}$, that is $\widehat{\lambda}_l = ((\|\boldsymbol{Y}_j\|_2)_{j=1,\dots,r})_{(r-l)}$.

4.2 Approximation of the mirror effect with group structure

Considering Definition 3.1 for the approximation of the mirror effect (Proposition 3.2), we find in the setting of group selection

$$\begin{aligned}
\widetilde{m}_l &= \frac{1}{n}\sum_{i=1}^{n} P(i \in \boldsymbol{S}; \boldsymbol{\mu}^{[-i]}) \left[E(\varepsilon_i^2 | i \in \boldsymbol{S}; \boldsymbol{\mu}^{[-i]}) - \sigma^2 \right] \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} P(\boldsymbol{S}_j = \mathbf{1}_j; \boldsymbol{\mu}^{[-j]}) \left[E(w^{-1} \left\| \boldsymbol{\varepsilon}_j \sigma^{-1} \right\|_2^2 | \boldsymbol{S}_j = \mathbf{1}_j; \boldsymbol{\mu}^{[-j]}) - 1 \right], \qquad (16)
\end{aligned}$$

where $\boldsymbol{\mu}^{[-j]}$ is the vector obtained by replacing $\boldsymbol{\mu}_j$ in $\boldsymbol{\mu}$ by zeros.

Following the approach from Section 3.2, the selection of the jth group can be expressed as $P(\boldsymbol{S}_j = \mathbf{1}_j; \boldsymbol{\mu}) = P(\|\boldsymbol{Y}_j\|_2 \geqslant T_{l,j})$ and, similarly,

$$E(w^{-1} \left\| \boldsymbol{\varepsilon}_j \sigma^{-1} \right\|_2^2 | \boldsymbol{S}_j = \mathbf{1}_j; \boldsymbol{\mu}) = E(w^{-1} \left\| \boldsymbol{\varepsilon}_j \sigma^{-1} \right\|_2^2 | \, \|\boldsymbol{Y}_j\|_2 \geqslant T_{l,j}),$$

where $T_{l,j}$ is a random threshold whose randomness comes from the ordering of the group norms, thus inducing dependence on the other groups $\boldsymbol{Y}_{j'}$ with $j' \neq j$. Given $\boldsymbol{\mu}^{[-j]}$, we can then write (16) as

$$\begin{aligned}
\widetilde{m}_l &= \frac{\sigma^2}{r}\sum_{j=1}^{r} P(\|\boldsymbol{\varepsilon}_j\|_2 \geqslant T_{l,j}) \left[E\left(\frac{1}{w} \|\boldsymbol{\varepsilon}_j/\sigma\|_2^2 \middle| \|\boldsymbol{\varepsilon}_j\|_2 \geqslant T_{l,j} \right) - 1 \right] \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} P\left(\|\boldsymbol{\varepsilon}_j/\sigma\|_2^2 \geqslant \frac{T_{l,j}^2}{\sigma^2} \right) \left[E\left(\frac{1}{w} \|\boldsymbol{\varepsilon}_j/\sigma\|_2^2 \middle| \|\boldsymbol{\varepsilon}_j/\sigma\|_2^2 \geqslant \frac{T_{l,j}^2}{\sigma^2} \right) - 1 \right] \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\int_{T_{l,j}^2/\sigma^2}^{\infty} \left(\frac{u}{w} - 1 \right) f_{\|\boldsymbol{\varepsilon}_j/\sigma\|_2^2}(u) du \right).
\end{aligned}$$

In practice, the regularisation parameter of the group lasso procedure acting as empirical threshold, $T_{l,j} = \widehat{\lambda}_l$, the approximated correction term $\widetilde{m}_l$ in the selection of l groups of variables

is estimated unbiasedly by

$$\widehat{m}_l = \sigma^2 \int_{\widehat{\lambda}_l^2/\sigma^2}^{\infty} (uw^{-1} - 1) f_{\|\boldsymbol{\varepsilon}_j/\sigma\|_2^2}(u) du.$$

This estimator can be plugged in into the modified information criterion of (13).

4.3 Signal-plus-noise model with Gaussian errors

Let $\boldsymbol{\varepsilon}$ be a n-vector of independent and identically distributed errors such that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. For a group of errors $\boldsymbol{\varepsilon}_j$, $\left\|\boldsymbol{\varepsilon}_j \sigma^{-1}\right\|_2^2 \sim \chi_w^2$ where χ_w^2 is a Chi-squared distribution with w degrees of freedom. Given the fact that $\int_0^{\infty} (uw^{-1} - 1) f_{\chi_w^2}(u) du = 0$, the mirror effect reduces to

$$\begin{aligned} \widetilde{m}_l &= \frac{\sigma^2}{r} \sum_{j=1}^{r} E\left(\int_0^{T_{l,j}^2/\sigma^2} (1 - uw^{-1}) f_{\chi_w^2}(u) du \right) \\ &= \frac{\sigma^2}{r} \sum_{j=1}^{r} E\left(F_{\chi_w^2}(T_{l,j}^2/\sigma^2) - F_{\chi_{w+2}^2}(T_{l,j}^2/\sigma^2) \right) \end{aligned} \tag{17}$$

with $F_{\chi_w^2}$ and $f_{\chi_w^2}$ the cumulative distribution function and density of the χ_w^2 distribution. When the group size w is 1 (singletons), Equation (17) reduces to the result found in [17]. Indeed,

$$\begin{aligned} \widetilde{m}_k = \widetilde{m}_l &= \frac{\sigma^2}{r} \sum_{j=1}^{r} E\left(F_{\chi_1^2}(T_{l,j}^2/\sigma^2) - F_{\chi_3^2}(T_{l,j}^2/\sigma^2) \right) \\ &= 2\sigma^2 n^{-1} \sum_{i=1}^{n} E\left(T_{k,i} \phi_\sigma(T_{k,i}) \right) \end{aligned} \tag{18}$$

ϕ_σ being the density of a zero-mean normal random variable with variance σ^2. Details for Equations (17) and (18) are given in Appendix B.

From Equation (17), we can see that, as the group size increases, $F_{\chi_w^2}$ and $F_{\chi_{w+2}^2}$ get closer, hence the mirror effect becomes smaller.

5 The mirror correction beyond the signal-plus-noise setting

When the design matrix in the regression model of (1) is not square and orthogonal, the problem cannot be transformed into a signal-plus-noise model. Although the full theoretical treatment of this general case lies beyond the scope of this paper, it is possible to find an expression for the mirror correction.

First of all, Assumption (A4) needs to be reformulated.

(A4ext) Given the binary vector $\boldsymbol{S}_{[-j]}$ containing all but the jth component of $\boldsymbol{S}$, we suppose that

$$S_j = 1 \Leftrightarrow \left| \boldsymbol{X}_j^T (\boldsymbol{Y} - \mathbf{X}_{[\boldsymbol{S}-j]} \widehat{\boldsymbol{\beta}}_{[\boldsymbol{S}-j]}) \right| \geqslant T_{\boldsymbol{S},j}, \tag{19}$$

where $\boldsymbol{X}_j$ is the jth column of the design matrix $\mathbf{X}$, and $\widehat{\boldsymbol{\beta}}_{[\boldsymbol{S}-j]}$ is the estimator defined by the selection in $\boldsymbol{S}_{[-j]}$ and $S_j = 0$. The random threshold $T_{\boldsymbol{S},j}$ is a function of $\left|\boldsymbol{X}_l^T(\boldsymbol{Y} - \mathbf{X}_{[\boldsymbol{S}-j]}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{S}-j]})\right|$, for all $l \neq j$.

The reasoning behind this assumption is that a component in $\boldsymbol{S}$ is active if the inner product of $\boldsymbol{X}_j$ with the residual of the estimator without that component is larger than a random threshold. In other words, if a component j is selected, then it would also be selected if the absolute inner product in (19) were larger, keeping the other inner products fixed. The inner product in (19) measures how much of the residual could be explained by the jth covariate.

As the mirror effect concerns the prediction error of least squares projections onto the selection $\boldsymbol{S}$, we have $\mathbf{X}_{\boldsymbol{S}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}} = \mathbf{P}_{\boldsymbol{S}}\boldsymbol{Y}$, with $\mathbf{P}_{\boldsymbol{S}}$ the orthogonal projection matrix as in (6). Thanks to the orthogonality, we have $\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{Y}\|_2^2 = \|\mathbf{P}_{[\boldsymbol{S}-j]}\boldsymbol{Y}\|_2^2 + \|(\mathbf{P}_{\boldsymbol{S}} - \mathbf{P}_{[\boldsymbol{S}-j]})\boldsymbol{Y}\|_2^2$, and furthermore

$$\boldsymbol{X}_j^T(\boldsymbol{Y} - \mathbf{X}_{[\boldsymbol{S}-j]}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{S}-j]}) = \boldsymbol{X}_j^T(\mathbf{X}_{\boldsymbol{S}}\widehat{\boldsymbol{\beta}}_{\boldsymbol{S}} - \mathbf{X}_{[\boldsymbol{S}-j]}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{S}-j]}) = \boldsymbol{X}_j^T((\mathbf{P}_{\boldsymbol{S}} - \mathbf{P}_{[\boldsymbol{S}-j]})\boldsymbol{Y}).$$

As a consequence, the assumption in (19) amounts to

$$\left\|(\mathbf{P}_{\boldsymbol{S}} - \mathbf{P}_{[\boldsymbol{S}-j]})\boldsymbol{Y}\right\|_2 \geqslant T'_{\boldsymbol{S},j},$$

where

$$T'_{\boldsymbol{S},j} = T_{\boldsymbol{S},j} / \left[\left\|\boldsymbol{X}_j^T\right\| \cos(\alpha_j)\right],$$

and α_j the angle between the vector $\boldsymbol{X}_j$ and the hyperplane $\mathbf{X}_{[\boldsymbol{S}-j]}$.

Let $\boldsymbol{S}^l, l = 0, 1, \ldots, k$ be a sequence of nested models with $\boldsymbol{S}_k = \boldsymbol{S}$, then

$$E\left[\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{\varepsilon}\|_2^2\right] = \sum_{l=1}^{k} E\left[\|(\mathbf{P}_{\boldsymbol{S}^l} - \mathbf{P}_{\boldsymbol{S}^{l-1}})\boldsymbol{\varepsilon}\|_2^2\right].$$

If the nested models were fixed independently from the observations, then in a normal, homoscedastic model, all values $\|(\mathbf{P}_{\boldsymbol{S}^l} - \mathbf{P}_{\boldsymbol{S}^{l-1}})\boldsymbol{\varepsilon}\|_2^2$ would be independent χ_1^2 distributed, just like in the orthogonal or identical design case. Denoting $\boldsymbol{\Delta}_l = \mathbf{P}_{\boldsymbol{S}^l} - \mathbf{P}_{\boldsymbol{S}^{l-1}}$, the mirror effect becomes

$$m_k = \frac{1}{n}\sum_{l=1}^{k} E\left[\|\boldsymbol{\Delta}_l\boldsymbol{\varepsilon}\|_2^2\right] - \frac{k}{n}\sigma^2.$$

A first step in finding an estimator of m_k consists of conditioning the expected value on the observed outcome of the variable selection, $\boldsymbol{S}$. The result is a random variable, depending on $\boldsymbol{S}$ and on the unobserved noise ε, making it as such unsuitable as estimator,

$$\begin{aligned}
\breve{m}_k &= \frac{1}{n}E\left[\|\mathbf{P}_{\boldsymbol{S}}\boldsymbol{\varepsilon}\|_2^2 \middle| \boldsymbol{S}; \boldsymbol{\beta}\right] - \frac{k}{n}\sigma^2. \qquad (20)\\
&= \frac{1}{n}\sum_{l=1}^{k} E\left[\|\boldsymbol{\Delta}_l\boldsymbol{\varepsilon}\|_2^2 \middle| \boldsymbol{S}_k; \boldsymbol{\beta}\right] - \frac{k}{n}\sigma^2\\
&= \frac{1}{n}\sum_{l=1}^{k} E\left[\|\boldsymbol{\Delta}_l\boldsymbol{\varepsilon}\|_2^2 \middle| \boldsymbol{S}_k; \|\boldsymbol{\Delta}_l(\boldsymbol{\mu} + \boldsymbol{\varepsilon})\|_2 \geqslant T'_{\boldsymbol{S}_k,l}\right] - \frac{k}{n}\sigma^2.
\end{aligned}$$

The last line should be read as conditioning the square of a noise component on the event that the component with response $\boldsymbol{\Delta}_l\boldsymbol{\mu}$ exceeds a threshold. The value of that threshold, a priori random, is fixed by the observation of the selection $\boldsymbol{S}_k$. As a result, the general design case reduces to the same framework as the signal-plus-noise case. Further development of the general case involves the use of the sparsity of $\boldsymbol{\Delta}_l\boldsymbol{\mu}$ in a way similar to the sparsity of $\boldsymbol{\mu}$ in the signal-plus-noise case. A second issue is the value of $T'_{\boldsymbol{S}_k,l}$. In the signal-plus-noise case, the thresholds are the same as the thresholds adopted in the lasso or group lasso estimators. In the case of orthogonal projection estimators, the value of threshold $T'_{\boldsymbol{S}_k,k}$ for the newly added covariate S_k is given by the value of $\|\boldsymbol{\Delta}_k\boldsymbol{Y}\|$, as this is the value that activates the selection of the covariate. In order to find the values of $T'_{\boldsymbol{S}_k,l}$ of previously activated covariates, $l < k$ that is, we need to identify the value of $\|\boldsymbol{\Delta}_l\boldsymbol{Y}\|$ so that the lth component, S_l, in $\boldsymbol{S}_k$ would be deactivated. As finding the exact value would be computationally complex, approximations can be constructed based on the actual value of the lasso parameter, $\widehat{\lambda}_k$, the value $\widehat{\lambda}_l$ that activated $\boldsymbol{S}_l$, and the value $T'_{\boldsymbol{S}^l,l} = \|\boldsymbol{\Delta}_l\boldsymbol{Y}\|$:

$$\widehat{\lambda}_{k,l} = T'_{\boldsymbol{S}^l,l}\widehat{\lambda}_k/\widehat{\lambda}_l,$$

leading to the following estimator of the mirror effect

$$\widehat{m}_k \approx \frac{1}{n}\sum_{l=1}^{k}\int_{\widehat{\lambda}_{k,l}}^{\infty}(u^2-\sigma^2)f_{\boldsymbol{\Delta}_l\boldsymbol{\varepsilon}}(u)du,$$

which can be plugged in into (13).

6 Illustration and simulation study

We generate $r = 200$ groups containing $w = 10$ coefficients $\boldsymbol{\mu}_j$ so that $\boldsymbol{\mu} = (\boldsymbol{\mu}_j)_{j=1,\dots,r}$ is a n-dimensional vector with $n = rw$. In the first instance, the simulation is designed to create the perfect setting for group selection by imposing that all components of $\boldsymbol{\mu}_j$ within group j are either all zero or all non-zero. A fraction $p = 1/20$ of the groups are non-zero, meaning that the total number of non-zeros in $\boldsymbol{\mu}$ equals exactly $n_1 = pn = prw = 100$. In the second instance, a proportion $q = 1/5$ of the n_1 non-zeros is transferred to the groups of zeros, in exchange for the same number of zeros, keeping the total number of non-zeros at n_1. The objective is to identify these n_1 components and to estimate their values.

The non-zeros $\boldsymbol{\mu}$ are distributed according to the zero inflated Laplace model $f_{\mu|\mu\neq 0}(\mu) = (a/2)\exp(-a|\mu|)$ where $a = 1/5$. The observations then are $(\boldsymbol{Y}_j)_{j=1,\dots,r} = (\boldsymbol{\mu}_j)_{j=1,\dots,r} + (\boldsymbol{\varepsilon}_j)_{j=1,\dots,r}$, where $\boldsymbol{\varepsilon}$ is a n-vector of independent, standard normal errors.

The simulation compares group selection using the criterion proposed in this paper with unstructured and group selection using classical Mallows' Cp.

Figure 1 has an illustration of the proposed variable selection on one sample of $n = 2000$ observations, $n_1 = 100$ of which have a non-zero value in $\boldsymbol{\mu}$. Among these, $(1-q)n_1 = 80$ appear in groups, while the other $qn_1 = 20$ are isolated. The estimations of the mirror effect corrections, plotted in Figure 1 (right), are calculated from equation (20) and their approximations from equations (17) and (18) with the empirical thresholds taken as the ordered group norms and the ordered absolute values of $\boldsymbol{Y}$ for group and unstructured selections respectively.

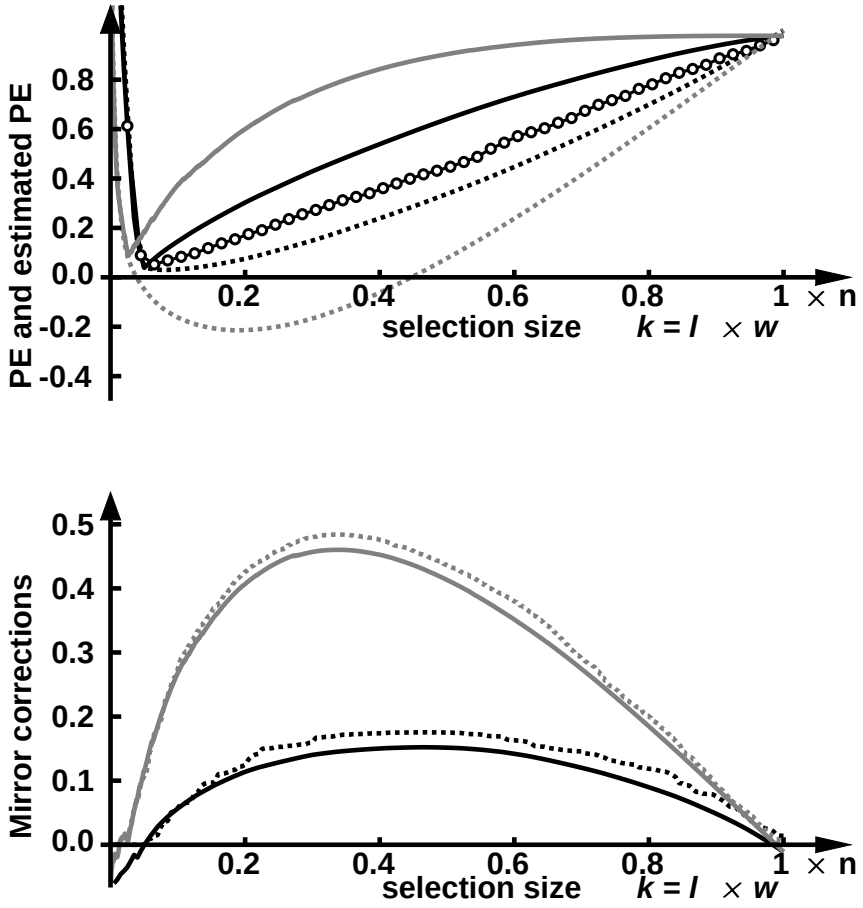

Figure 1: (top) The gray and black solid lines represent the PE for unstructured and group selections and the dashed lines represent their respective Mallows' Cp in the signal-plus-noise model. The black bullet line depicts the oracular mirror, $\mathrm{PE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{O}_k})$ as in (4), for the group selection. The mirror for the unstructured selection is not depicted for the sake of clarity. It almost coincides with the mirror for the group selection. (bottom) The mirror effects estimated with Equation (20) for unstructured and group selections are plotted in gray and black solid lines and their approximations from Equations (18) and (17) in gray and black dotted lines. The samples size n equals 2000 in this case.

		Unstructured (best k)			Group selection		
		Cp	Cp+corr.	PE	Cp	Cp+corr	PE
k	Q_1	365	53	53	140	100	100
	Q_2	375	58	57	150	100	100
	Q_3	387	63	61	170	100	100
	mean	375.327	58.448	57.263	155.66	102.65	100
PE	Q_1	0.576	0.112	0.102	0.095	0.047	0.045
	Q_2	0.597	0.124	0.113	0.112	0.052	0.050
	Q_3	0.622	0.139	0.124	0.133	0.059	0.054
	mean	0.599	0.126	0.113	0.114	0.054	0.050
FP	Q_1	300	12	11	60	20	20
	Q_2	310	14	13	70	20	20
	Q_3	321.5	17	15	90	20	20
	mean	310.734	14.687	13.168	75.052	22.625	20
FN	Q_1	33	53	53	19	20	20
	Q_2	35	56	56	20	20	20
	Q_3	38	60	59	20	20	20
	mean	35.407	56.239	55.905	19.392	19.975	20

Table 1: Results of simulation study for 1000 samples following in the same settings as the illustration in Figure 1. Each sample has exactly $n_1 = 100$ non-zero values in the vector $\boldsymbol{\mu}$ of length $n = 2000$. Among the $n_1 = 100$ non-zeros, $qn_1 = 20$ do not belong to a group, while the groups of non-zeros actually contain $qn_1 = 20$ zero values in $\boldsymbol{\mu}$. The table lists the lower quartile Q_1, median Q_2, upper quartile Q_3 and mean values of four quantities, the selection size k, the prediction error PE, the absolute number of false positives FP, and the number of false negatives, FN. Selection is performed by simple minimisation of Cp, minimisation of Cp, corrected as in [17], minimisation of the prediction error (as a benchmark), and with the same three finetuning approaches applied to group selection. See text for a discussion of the results.

Figure 1 (left) plots the prediction error and Mallows' Cp as functions of the model size for unstructured and group selections. In each case, we observe that the PE and Cp curves are reflexion of each other with respect to a mirror curve. It can be seen that the minimum PE in the case of group selection is lower than the minimum PE in the case of unstructured selection. That means that with 80% of the non-zero's occurring in groups, it is beneficial to apply group selection. Also, the selection size at the minimum is larger in the group selection. These conclusions are confirmed in the simulation study of Table 1, summarising the results of 1000 samples. The fifth column in the table corresponds to the approach presented in this paper. It shows a near-perfect balance between false negatives and false positives, with virtually in every simulation run 20 false negatives and 20 false positives. This is the best that can be expected from a procedure that selects components in groups. Isolated non-zeros cannot be detected, while zeros within the group cannot be excluded, unless a post-processing is applied. Simulations with a proportion $q = 0$ of isolated non-zeros (not tabled here) have been run as well, leading to zero false positives and zero false negatives in virtually all runs. Software

reproducing these results is available.

The prediction error of the data driven corrected information criterion is not far from the minimum prediction error in the sixth column (obtained by an oracle knowing for each candidate model the exact prediction error). The simulation study also confirms the observation in Figure 1 that in the setting of this simulation, the group selection reduces the prediction error. In particular, structuring the selection helps in finding less prominent non-zeros, at least when they appear in a group, thus reducing the number of false negatives. In order to pick up some of the remaining isolated non-zeros, a post-processing performing a sort of model averaging is still to be investigated.

7 Application to image denoising

In this section, we compare the performance of unstructured and group selections for noise reduction, using both Mallows' Cp and its mirror-corrected counterpart as a quality rule. We apply these methods to the MNIST dataset from [21], consisting of 28×28 pixel greyscale images of handwritten digits. The main idea here is that images of a same digit share common features, hence we find a structure across these images. Sparse representations are obtained by application of a Haar transform on the images. Writing the images with noise by $\boldsymbol{Y} = \boldsymbol{\mu} + \sigma \boldsymbol{Z}$, the Haar basis acts as the design matrix $\mathbf{X}$ in the sparse regression model of (1), where $\boldsymbol{\beta}$ is the vector of Haar coefficients. In this application, the forward Haar transform corresponds to $\mathbf{X}^{-1}$, leading to the sparse signal-plus-noise model in the transformed domain $\mathbf{X}^{-1}\boldsymbol{Y} = \boldsymbol{\beta} + \sigma \mathbf{X}^{-1}\boldsymbol{Z}$. Using structured selection should therefore improve denoising. The technique could be used with pictures from a surveillance camera where the background is fixed. This way, one could detect changes such as unexpected objects in relatively limited areas of the pictures, or perhaps recover the original background more precisely if the pictures are noisy. In the next paragraphs, we describe the methodology for preparing the MNIST dataset and the selection procedure before presenting our results.

7.1 Methodology

For a given digit, 10 corresponding images are randomly chosen from the training set of the MNIST database. The values of the pixels are adjusted so that the images contain their values in the range 0 (black) to 1 (white). The black background is extended with a 2-pixels border to obtain 32×32 pixel images. Finally a Gaussian white noise with variance 0.04 is added to the images.

We apply simple Haar wavelets on each of the 10 noisy images and perform selection on the detail coefficients. Both unstructured and group selections are estimated: as the images are similar, it is reasonable to assume that if a detail is non-zero for one image, then the same details from the other images should also be non-zeros, hence the grouping. Following this idea, we get 32×32 groups of 10 coefficients. The best models are found by optimising the Mallows' Cp criterion (3) and the mirror-corrected Cp using Equations (17) and (18), then we recover the corresponding images which we compare to the original noise-free images.

Table 2: Summary of average performances for the models optimising the PE, Mallows' Cp and mirror-corrected Cp in unstructured and group selections.

	Unstructured selection			Group selection		
	PE	Cp	Cp+2m_k	PE	Cp	Cp+2m_l
Sum of residuals	0.0136	0.0268	0.0139	0.0113	0.0130	0.0115
# non-zeros details	591	2336	575	1459	2060	1480
# false positives	45	1291	43	388	782	404
# misspecified values	1527	2273	1537	1344	1532	1356
% true neg. recovery	99.45	84.28	99.48	95.28	90.47	95.08
% true pos. recovery	26.92	51.57	26.28	52.83	63.03	53.07

These steps are repeated 1000 times, which allows us to compute an average value of the difference between noise-reduced images and their respective noise-free images.

7.2 Results

We choose to focus on images of the digit "7". A summary of diverse results for the models optimising the PE, Mallows' Cp and mirror-corrected Cp in unstructured and group selections is listed in Table 2. In both configurations, correcting the Mallows' Cp criterion with the mirror effect leads to models close to those one would get optimising the prediction error. Note that the minimum prediction error serves as a benchmark in this comparison. Indeed, without knowledge of the true underlying data, it cannot be computed exactly. Hence, the fact that the mirror corrected Cp method comes close to the minimum prediction error method illustrates the good performance of the method.

Amongst the 10240 details calculated from the 10 noise-free images, an average of 2027.6 are non-zeros. Under unstructured selection, the minimum mirror-corrected Cp model underestimates the true number of non-zero details: although it has the lowest number of false positives, it also has a lower percentage of true positive recovery than the minimum PE and minimum uncorrected Cp selections. The minimum uncorrected Cp model contains the largest numbers of false positives and misspecified values (false positives and false negatives); this can be partly explained by the overestimation of its model size.

Under group selection, we observe a similar behaviour but less extreme: in particular, the best model using the mirror contains more false positives compared to the unstructured case, but its percentage of true positive recovery is twice as large. In terms of sum of residuals, Mallows' Cp under unstructured selection performs poorly (0.02678), whereas the smallest error is achieved using the corrected Cp under group selection (0.01150). The errors of unstructured corrected Cp and group uncorrected Cp are 0.01385 and 0.01304 respectively.

A sample of 5 images from one iteration is available in Appendix C.

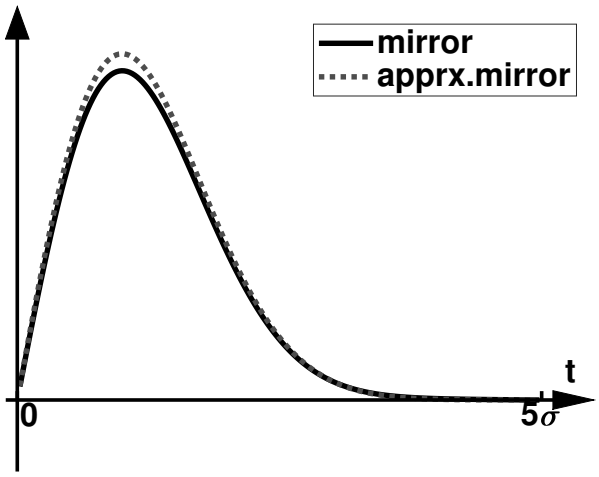

Figure 2: Plot of $(-1/n)\sum_{i=1}^{n} h(t;\mu_i)$ as a function of t, along with an approximation $h(t;0)$.

8 Conclusion

In this paper, we presented a new approximation of the degrees of freedom for use in a Mallows' Cp driven group Lasso selection of sparse covariates in a high-dimensional model. The approximation develops the so-called mirror effect, which compensates for the effect of false positives on the optimisation of the information criterion. This does not mean that the proposed method prevents all false positives from occuring. Indeed, it has been found that false positives happen early on in the Lasso selection process [**?**]. On the other hand, the group structured selection reduces the number of isolated false positives. The mirror correction on top of the structured selection keeps the additional false positives due to the optimisation process under control. The paper has found an explicit expression for the mirror correction in the setting of group Lasso with additive, normal uncorrelated noise.

Appendices

A Proof of Proposition 3.2

Proof.

The key point in the proof is to realise that the largest contributions to the approximation error come from the values in μ away from zero. Using the assumption of asymptotic sparsity, these contributions become less and less important.

Defining

$$\bar{h}(t;\boldsymbol{\mu}) = \frac{1}{n}\sum_{i=1}^{n} h(t;\mu_i),$$

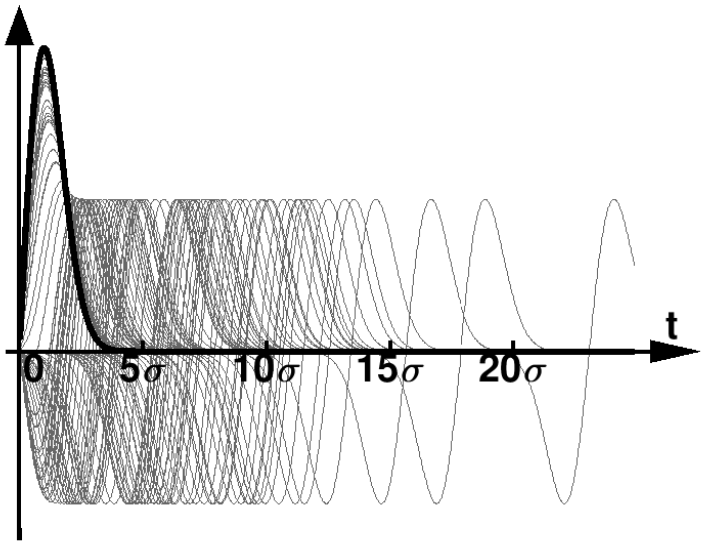

Figure 3: Plots of $h(t;\mu_i)$ with in bold line, $h(t;0)$.

we can write from (9),

$$\begin{aligned} m_k - \widetilde{m}_k &= E\left[\bar{h}(T_{k,i};\boldsymbol{\mu}) - \bar{h}(T_{k,i};\mathbf{0})\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[h(T_{k,i};\mu_i) - h(T_{k,i};0)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[2G_\varepsilon(T_{k,i}) - G_\varepsilon(T_{k,i}-\mu_i) - G_\varepsilon(T_{k,i}+\mu_i)\right]. \end{aligned}$$

The value of $\bar{h}(t;\boldsymbol{\mu})$ is depicted as a function of t in Figure 2. The individual contributions $h(t;\mu_i)$ for a typical sparse signal are plotted in Figure 3.

We construct an upper bound for $h(t;\mu)$, consisting of three parts, depending on the value of μ. First, we have a general upper bound

$$|h(t;\mu) - h(t;0)| \leqslant \max_t |2G_\varepsilon(t) - G_\varepsilon(t-\mu) - G_\varepsilon(t+\mu)| \leqslant 4\max_x |G_\varepsilon(x)| = 4|G_\varepsilon(\sigma)|,$$

as indeed, on the positive axis, $|G_\varepsilon(x)|$ is unimodal with global maximum in $x=\sigma$.

Remark A.1 *The upper bound is pessimistic, since* $\lim_{t\to\infty}|G_\varepsilon(t)| = 0$*, so for every* $\eta > 0$*, there exists a* t^**, so that for* $t > t^*$*, we find* $2|G_\varepsilon(t) - G_\varepsilon(t+\mu_i)| \leqslant |2G_\varepsilon(t)| < 2\eta$*, and so* $|2G_\varepsilon(t) - G_\varepsilon(t-\mu_i) - G_\varepsilon(t+\mu_i)| < 2|G_\varepsilon(\sigma)| + 2\eta$.

The second part of the upper bound is for small values of μ, as illustrated in Figure 4. Let

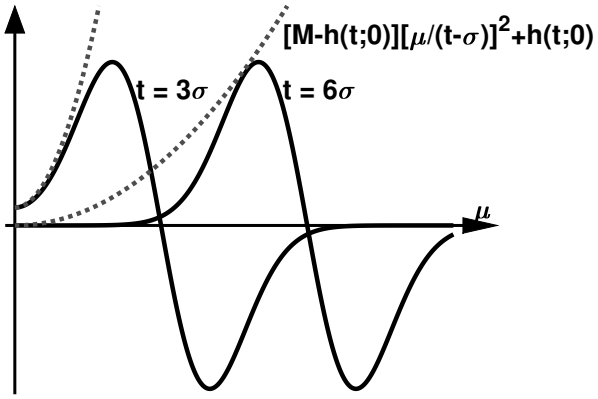

Figure 4: In bold line, plots of $h(t;\mu_i)$ as a function of μ_i for two values of t. The dotted lines represent the upper bounds.

M be a constant, a priori depending on t, so that for $|\mu| \leqslant t-\sigma$, we have that

$$h(t;\mu) - h(t;0) \leqslant [M - h(t;0)] \cdot \left[\frac{\mu}{t-\sigma}\right]^2. \tag{21}$$

This construction is possible since $h'(t;0) = 0$.

Remark A.2 *Obviously, one can take* $[M - h(t;0)]/(t-\sigma)^2$ *to be equal to* $\max_{x\in\mathbb{R}} 2|G''_\varepsilon(x)|$, *but that choice would lead to a pessimistic upper bound when* t *grows larger. We will keep* $2\max_{x\in\mathbb{R}}|G''_\varepsilon(x)|$ *as an upper bound when* $[M - h(t;0)]/(t-\sigma)^2$ *is replaced by the random version* $[M - h(t;0)]/(T_{k,i}-\sigma)^2$.

For $t \geqslant \sigma$ and $\mu \geqslant \tau$, we have that $-G_\varepsilon(t+\tau+\mu) \leqslant -G_\varepsilon(t-\tau+\mu)$, and so that $h(t+\tau;\mu) \leqslant h(t;\mu-\tau)$, and thus

$$h(t+\tau;\mu) \leqslant h(t;0) + [M - h(t;0)] \cdot \left[\frac{\mu-\tau}{t-\sigma}\right]^2.$$

For t sufficiently large, $h(t;0) - h(t+\tau;0)$ is small enough for any τ, so that

$$\frac{h(t;0) - h(t+\tau;0)}{M - h(t+\tau;0)} \leqslant \left(\frac{\tau}{t+\tau-\sigma}\right)^2.$$

This is equivalent to

$$h(t;0) \leqslant h(t+\tau;0) + [M - h(t+\tau;0)] \cdot \left[\frac{\tau}{t+\tau-\sigma}\right]^2. \tag{22}$$

This implies that on $[\tau, t+\tau-\sigma]$,

$$h(t;0) + [M - h(t+\tau;0)] \cdot \left[\frac{\mu-\tau}{t-\sigma}\right]^2 \leqslant h(t+\tau;0) + [M - h(t+\tau;0)] \cdot \left[\frac{\mu}{t+\tau-\sigma}\right]^2 \tag{23}$$

as indeed both quadratic forms have the same value, M, at $\mu = t + \tau - \sigma$, while for $\mu = \tau$ this reduces to (22). The right hand side of (23) has the same form as the right hand side in (21). As a result, for t sufficiently large, the constant M in (21) does not depend on t. By choosing a value for M larger than $4|G_\varepsilon(\sigma)|$, the upper bound in (21) holds for any μ. Taking into account Remark A.2, we can write

$$h(t;\mu) - h(t;0) \leqslant q(t)\mu^2,$$

where

$$q(t) = \min\left(2\max_{x\in\mathbb{R}}|G''_\varepsilon(x)|, \frac{M - h(t;0)}{(t-\sigma)^2}\right).$$

For large values of μ, we however need a third and tighter upper bound. Because of the symmetry in $f_\varepsilon(x)$, we have for $\mu = 2t$, that $h(2t;t) = -G_\varepsilon(3t) - G_\varepsilon(-t) = -G_\varepsilon(3t) + G_\varepsilon(t)$ and as $2t$ is far beyond the largest local maximum of $|h(\mu;t)|$ as a function of μ, it holds for $\mu > 2t$ that $h(\mu;t) > h(2t;t)$, and so

$$|h(\mu;t) - h(0;t)| = h(0;t) - h(\mu;t) < h(0;t) - h(2t;t) = |3G_\varepsilon(t) - G_\varepsilon(3t)| < 3|G_\varepsilon(t)|.$$

The three parts of the analysis allow us to conclude that the approximation error of $\widetilde{m}_k$ is bounded by

$$\begin{aligned}
|m_k - \widetilde{m}_k| &\leqslant \frac{1}{n}\sum_{i=1}^{n} && E\left|h(T_{k,i};\mu_i) - h(T_{k,i};0)\right| \\
&\leqslant \frac{1}{n}\sum_{i=1}^{n} && P(T_{k,i} > |\mu_i|/2)E\left(q(T_{k,i})|T_{k,i} > |\mu_i|/2\right)\mu_i^2 \\
& && +P(T_{k,i} \leqslant |\mu_i|/2)3E\left(|G_\varepsilon(T_{k,i})|\,|T_{k,i} \leqslant |\mu_i|/2\right). \qquad (24)
\end{aligned}$$

Moreover, it is easy to find a constant K so that $q(t) \leqslant K/t$, for all values of t, and also, because $q(t)$ is a monotonously non-increasing function, we have

$$E\left(q(T_{k,i})|T_{k,i} > |\mu_i|/2\right) \leqslant E\left(q(T_{k,i})\right) \leqslant KE\left(1/T_{k,i}\right) \to 0 \text{ when } n \to \infty.$$

Combining this with Assumption (11), we find for the first sum in (24),

$$\frac{1}{n}\sum_{i=1}^{n} P(T_{k,i} > |\mu_i|/2)E\left(q(T_{k,i})|T_{k,i} > |\mu_i|/2\right)\mu_i^2 = o\left[\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right].$$

For the second sum in (24), we see that if $P(T_{k,i} \leqslant |\mu_i|/2)$ does not tend to zero, then by Markov's inequality, we have

$$P\left(\frac{1}{T_{k,i}} > \frac{2}{|\mu_i|}\right) \leqslant E\left(\frac{1}{T_{k,i}}\right)\frac{|\mu_i|}{2} \Rightarrow |\mu_i| \geqslant \frac{2P\left(\frac{1}{T_{k,i}} > \frac{2}{|\mu_i|}\right)}{E\left(\frac{1}{T_{k,i}}\right)} \to \infty.$$

With $|\mu_i| \to \infty$ and $E\left(1/T_{k,i}\right) \to 0$, the value of $E\left(|G_\varepsilon(T_{k,i})| \mid T_{k,i} \leqslant |\mu_i|/2\right)$ then tends to $E\left(|G_\varepsilon(T_{k,i})|\right)$ which in turn tends to zero since, under Assumption (A2), there exists a constant L so that $|G_\varepsilon(t)| \leqslant L/t$. As a result, we have

$$\frac{1}{n}\sum_{i=1}^{n} P(T_{k,i} \leqslant |\mu_i|/2)3E\left(|G_\varepsilon(T_{k,i})||T_{k,i} \leqslant |\mu_i|/2\right) = o\left[\mathrm{PE}(\widehat{\boldsymbol{\mu}}_k)\right],$$

thereby completing the proof. □

B Development of calculations for Equations 17 and 18

With Γ the gamma function and $F_{\chi^2_w}$ and $f_{\chi^2_w}$ the cumulative distribution function and density of the χ^2_w distribution, we find the following result for Equation (17):

$$\begin{aligned}
\widetilde{m}_l &= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\int_0^{\frac{T^2_{l,j}}{\sigma^2}} (1-uw^{-1}) f_{\chi^2_w}(u)du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\int_0^{\frac{T^2_{l,j}}{\sigma^2}} (1-uw^{-1}) \frac{1}{2^{\frac{w}{2}}\Gamma(\frac{w}{2})} u^{\frac{w}{2}-1}e^{-\frac{u}{2}}du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\int_0^{\frac{T^2_{l,j}}{\sigma^2}} \frac{1}{2^{\frac{w}{2}}\Gamma(\frac{w}{2})} u^{\frac{w}{2}-1}e^{-\frac{u}{2}}du - \frac{1}{w}\int_0^{\frac{T^2_{l,j}}{\sigma^2}} u\frac{1}{2^{\frac{w}{2}}\Gamma(\frac{w}{2})} u^{\frac{w}{2}-1}e^{-\frac{u}{2}}du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(F_{\chi^2_w}(T^2_{l,j}\sigma^{-2}) - \frac{2\Gamma(\frac{w}{2}+1)}{w\Gamma(\frac{w}{2})}\int_0^{\frac{T^2_{l,j}}{\sigma^2}} \frac{1}{2^{\frac{w}{2}+1}\Gamma(\frac{w}{2}+1)} u^{\frac{w}{2}}e^{-\frac{u}{2}}du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(F_{\chi^2_w}(T^2_{l,j}\sigma^{-2}) - \frac{2\Gamma(\frac{w}{2}+1)}{w\Gamma(\frac{w}{2})} F_{\chi^2_{w+2}}(T^2_{l,j}\sigma^{-2})\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(F_{\chi^2_w}(T^2_{l,j}\sigma^{-2}) - F_{\chi^2_{w+2}}(T^2_{l,j}\sigma^{-2})\right)
\end{aligned}$$

Also, when the group size w is 1 (singletons), Equation (17) reduces to Equation (18):

$$\begin{aligned}
\widetilde{m}_k &= \widetilde{m}_l = \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(F_{\chi^2_1}(T^2_{l,j}\sigma^{-2}) - F_{\chi^2_3}(T^2_{l,j}\sigma^{-2})\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\frac{1}{\Gamma(\frac{1}{2})}\int_0^{\frac{T^2_{l,j}}{2\sigma^2}} u^{-\frac{1}{2}}e^{-u}du - \frac{1}{\Gamma(\frac{3}{2})}\int_0^{\frac{T^2_{l,j}}{2\sigma^2}} u^{\frac{1}{2}}e^{-u}du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\frac{1}{\Gamma(\frac{1}{2})}\int_0^{\frac{T_{l,j}}{\sigma}} \left(\frac{u^2}{2}\right)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}udu - \frac{2}{\Gamma(\frac{1}{2})}\int_0^{\frac{T_{l,j}}{\sigma}} \left(\frac{u^2}{2}\right)^{\frac{1}{2}} e^{-\frac{u^2}{2}}udu\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\frac{1}{\Gamma(\frac{1}{2})}\int_0^{\frac{T_{l,j}}{\sigma}} \sqrt{2}e^{-\frac{u^2}{2}}du - \frac{2}{\Gamma(\frac{1}{2})}\int_0^{\frac{T_{l,j}}{\sigma}} \frac{1}{\sqrt{2}}u^2e^{-\frac{u^2}{2}}du\right) \\
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\frac{2}{\sqrt{2}\Gamma(\frac{1}{2})}\int_0^{\frac{T_{l,j}}{\sigma}} (1-u^2)e^{-\frac{u^2}{2}}du\right)
\end{aligned}$$

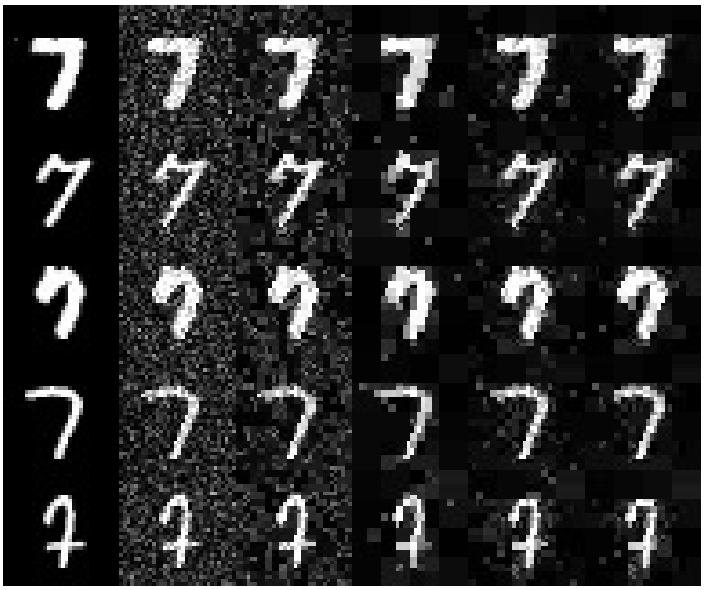

Figure 5: Image denoising for 10 handwritten digits using Mallows' Cp and the mirror-corrected Cp in unstructured and group selections.

$$
\begin{aligned}
&= \frac{\sigma^2}{r}\sum_{j=1}^{r} E\left(\frac{2}{\sqrt{2}\Gamma(\frac{1}{2})}\frac{T_{l,j}}{\sigma}e^{-\frac{T_{l,j}^2}{2\sigma^2}}\right)\\
&= 2\sigma^2 r^{-1}\sum_{j=1}^{r} E\left(T_{l,j}\phi_\sigma(T_{l,j})\right) = 2\sigma^2 n^{-1}\sum_{i=1}^{n} E\left(T_{k,i}\phi_\sigma(T_{k,i})\right)
\end{aligned}
$$

as $e^{-\frac{u^2}{2}} - u^2 e^{-\frac{u^2}{2}}$ is the derivative of $ue^{-\frac{u^2}{2}}$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, ϕ_σ being the density of a zero-mean normal random variable with variance σ^2.

C Illustration for image denoising

In its first column, Figure 5 presents a sample of 5 noise-free images, then the noisy ones in the second column. Finally the denoised images using Mallows' Cp and the mirror-corrected Cp in unstructured and group selections are shown in the third to fourth, and fifth to sixth columns respectively.

References

[1] Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: B. Petrov, F. Csáki (eds.) Second International Symposium on Information Theory, pp. 267–281. Akadémiai Kiadó, Budapest (1973)

[2] Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. Bernoulli **19**(2), 521–547 (2013)

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

[4] Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging, first edn. Cambridge University Press (2008)

[5] Das, D., Chatterjee, A., Lahiri, S.N.: Higher order refinements by bootstrap in lasso and other penalized regression methods. Tech. rep., Indian Institute of Technology, Kanpur; Indian Statistical Institute, Delhi; Washington University in St. Louis (2020). URL `https://arxiv.org/abs/1909.06649`

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

[7] Donoho, D.L.: De-noising by soft-thresholding. IEEE Transactions on Information Theory **41**(3), 613–627 (1995)

[8] Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Comm. on Pure and Appliead Mathematics **59**, 797–829 (2006)

[9] Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. American Statistical Association **90**(432), 1200–1224 (1995)

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

[11] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. American Statistical Association **96**(456), 1348–1360 (2001)

[12] Foygel Barber, R., Candès, E.: Controlling the false discovery rate via knockoffs. The Annals of Statistics **43**(5), 2055–2085 (2015)

[13] Friedman, J., Hastie, T., Hofling, H., Tibshirani, R.: Pathwise coordinate optimization. The Annals of Applied Statistics **1**(2), 302–332 (2007)

[14] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. Biostatistics **9**(3), 432–441 (2008)

[15] Fu, W.: Penalized regressions: the bridge vs the lasso. Journal of Computational and Graphical Statistics **7**(3), 397–416 (1998)

[16] Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high-dimensional models. Statistical science **27**(4), 481–499 (2012)

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

[18] Jansen, M.: Generalized cross validation in variable selection with and without shrinkage. Journal of Statistical Planning and Inference **159**, 90–104 (2015)

[19] Javanmard, A., Montanari, A.: Debiasing the lasso: Optimal sample size for gaussian designs. The Annals of Statistics **46**(6A), 2593–2622 (2018)

[20] Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Processes. Springer Series in Statistics. Springer, 175 Fifth Avenue, New York 10010, USA (1983)

[21] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

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

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

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

[25] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, Series B **67**(1), 91–108 (2005)

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

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

[28] Wainwright, M.J.: 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 (2009)

[29] Wang, H., Leng, C.: A note on adaptive group lasso. Computational Statistics and Data Analysis **52**(12), 5277–5286 (2008)

[30] Yang, Y.: Can the strengths of aic and bic be shared? Biometrika **92**, 937–950 (2005)

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

[32] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B **68**, 49–67 (2006)

[33] Zhang, C.: Nearly unbiased variable selection under the minimax concave penalty. The Annals of Statistics **38**(2), 894–942 (2010)

[34] Zhao, P., Rocha, G., Yu, B.: The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics **37**, 3468–3497 (2009)

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

[36] Zou, H.: The adaptive lasso and its oracle properties. J. American Statistical Association **101**, 1418–1429 (2006)

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