Beyond support in two-stage variable selection

Jean-Michel Bécu[1] · **Yves Grandvalet**[1] · **Christophe Ambroise**[2] · **Cyril Dalmasso**[2]

Received: 2 July 2015 / Accepted: 5 November 2015 / Published online: 20 November 2015

Abstract Numerous variable selection methods rely on a two-stage procedure, where a sparsity-inducing penalty is used in the first stage to predict the support, which is then conveyed to the second stage for estimation or inference purposes. In this framework, the first stage screens variables to find a set of possibly relevant variables and the second stage operates on this set of candidate variables, to improve estimation accuracy or to assess the uncertainty associated to the selection of variables. We advocate that more information can be conveyed from the first stage to the second one: we use the magnitude of the coefficients estimated in the first stage to define an adaptive penalty that is applied at the second stage. We give the example of an inference procedure that highly benefits from the proposed transfer of information. The procedure is precisely analyzed in a simple setting, and our large-scale experiments empirically demonstrate that actual benefits can be expected in much more general situations, with sensitivity gains ranging from 50 to 100 % compared to state-of-the-art.

✉ Yves Grandvalet
yves.grandvalet@utc.fr

Jean-Michel Bécu
jean-michel.becu@utc.fr

Christophe Ambroise
christophe.ambroise@genopole.cnrs.fr

Cyril Dalmasso
cyril.dalmasso@genopole.cnrs.fr

[1] Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc UMR 7253, CS 60 319, 60 203 Compiègne Cedex, France

[2] LaMME, Université d'Évry val d'Essonne, 23 Boulevard de France, 91000 Évry, France

Keywords Linear model · Lasso · Variable selection · p-values · False discovery rate · Screen and clean

1 Introduction

The selection of explanatory variables has attracted much attention these last two decades, particularly for high-dimensional data, where the number of variables is greater than the number of observations. This type of problem arises in a variety of domains, including image analysis (Wang et al. 2008), chemometry (Chong and Jun 2005) and genomics (Xing et al. 2001; Ambroise and McLachlan 2002; Anders and Huber 2010).

1.1 Motivations

Since the development of the sparse estimators derived from ℓ_1 penalties such as the Lasso (Tibshirani 1996) or the Dantzig selector (Candès and Tao 2007), sparse models have been shown to be able to recover the subset of relevant variables in various situations (see, e.g. Candès and Tao 2007; Verzelen 2012; Bühlmann 2013; Tenenhaus et al. 2014).

However, the conditions for support recovery are quite stringent and difficult to assess in practice. Furthermore, the strength of the penalty to be applied differs between the problem of model selection, targeting the recovery of the support of regression coefficients, and the problem of estimation, targeting the accuracy of these coefficients. As a result, numerous variable selection methods rely on a two-stage procedure, where the Lasso is used in the first stage to predict the support, which is then conveyed to the second stage for estimation or inference purposes. In this framework, the first stage screens variables to find a set of possibly relevant variables and the second stage operates on this set of candidate variables, to improve estimation accuracy or to assess the uncertainty associated to the selection of variables.

This strategy has been proposed to correct for the estimation bias of the Lasso coefficients, with several variants in the second stage. The latter may then be performed by ordinary least squares (OLS) regression for the LARS/OLS Hybrid of Efron et al. (2004) (see also Belloni and Chernozhukov 2013), by the Lasso for the Relaxed Lasso of Meinshausen (2007), by modified least squares or ridge regression for Liu and Yu (2013), or with "any reasonable regression method" for the marginal bridge of Huang et al. (2008).

The same strategy has been proposed to perform variable selection with statistical guarantees by Wasserman and Roeder (2009), whose approach was pursued by Meinshausen et al. (2009). The first stage performs variable selection by Lasso or other regression methods on a subset of data. It is followed by a second stage relying on the OLS, on the remaining subset of data, to test the relevance of these selected variables.[1]

To summarize, the first stage of these approaches screens variables and transfers the estimated support of variables to the second stage for a more focused in-depth analysis. In this paper, we advocate that more information can be conveyed from the first stage to the second one, by using the magnitude of the coefficients estimated in the first stage. Improving this information transfer is essential in the so-called large p small n designs which are typical in genomic applications. The magnitude of regression coefficients, which is routinely interpreted as a quantitative gauge of relevance in statistical analysis, can be used to define an adaptive penalty, following alternative views of sparsity-inducing penalties. These views may originate from variational methods regarding optimization, or from hierarchical Bayesian models, as detailed in Sect. 1.2. In Sect. 2, we give an example of procedure that highly benefits from the proposed transfer of magnitude. This benefit is thoroughly analyzed in Sect. 3 for the orthonormal setting. The actual benefits are empirically demonstrated in diverse situations in Sect. 4.

1.2 Beyond support: magnitude

We consider the following high-dimensional sparse linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^t$ is the vector of responses, $\mathbf{X}$ is the $n \times p$ design matrix with $p \gg n$, $\boldsymbol{\beta}^\star$ is the sparse p-dimensional vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is a n-dimensional vector of independent random variables of mean zero and variance σ^2.

We discuss here two-stage approaches relying on a first screening of variables based on the Lasso, which is nowadays widely used to tackle simultaneously variable estimation and selection.[2]

The original Lasso estimator is defined as:

$$\hat{\boldsymbol{\beta}}(\lambda) = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} J(\boldsymbol{\beta}) + \lambda \left\|\boldsymbol{\beta}\right\|_1 , \tag{1}$$

where λ is a hyper-parameter, and $J(\boldsymbol{\beta})$ is the data-fitting term. Throughout this paper, we will discuss regression problems for which $J(\boldsymbol{\beta})$ is defined as

$$J(\boldsymbol{\beta}) = \frac{1}{2} \left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right\|_2^2 ,$$

but, except for the numerical acceleration tricks mentioned in Appendix 2, the overall feature selection process may be applied to any other form of $J(\boldsymbol{\beta})$, thus allowing to address classification problems.

Our approach relies on an alternative view of the Lasso, seen as an adaptive-ℓ_2 penalization scheme, following a viewpoint that has been mostly taken for optimization purposes (Grandvalet 1998; Grandvalet and Canu 1999; Bach et al. 2012). This view is based on a variational form of the Lasso:

$$\begin{aligned} &\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\tau} \in \mathbb{R}^p} \quad J(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \frac{1}{\tau_j} \beta_j^2 \\ &\qquad \text{s. t.} \sum_{j=1}^{p} \tau_j - \sum_{j=1}^{p} \left|\beta_j\right| \leq 0 \\ &\qquad\qquad \tau_j \geq 0, \;\; j = 1, \ldots, p. \end{aligned} \tag{2}$$

The variable $\boldsymbol{\tau}$ introduced in this formulation, which adapts the ℓ_2 penalty to the data, can be shown to lead to the following adaptive-ridge penalty:

$$\sum_{j=1}^{p} \frac{\lambda}{|\hat{\beta}_j(\lambda)|} \beta_j^2 , \tag{3}$$

where the coefficients $\hat{\beta}_j(\lambda)$ are the solution to the Lasso problem (1).

[1] In their two-stage procedure, Liu and Yu (2013) also proposed to construct confidence regions and to conduct hypothesis testing by bootstrapping residuals. Their approach fundamentally differs from Wasserman and Roeder (2009), in that inference does not rely on the two-stage procedure itself, but on the properties of the estimator obtained in the second stage.

[2] Though many sparsity-inducing penalties, such as the Elastic-Net, the group-Lasso or the fused-Lasso lend themselves to the approach proposed here, the paper is restricted to the simple Lasso penalty.

Using this adaptive-ℓ_2 penalty returns the original Lasso estimator (see proof in Appendix 1). This equivalence is instrumental here for defining the data-dependent penalty (3), implicitly determined in the first stage through the Lasso estimate, that will also be applied in the second stage. In this process, our primary aim is to retain the magnitude of the coefficients of $\hat{\boldsymbol{\beta}}(\lambda)$ in addition to the support $\mathscr{S}_\lambda = \{j \in \{1, ..., p\} | \hat{\beta}_j(\lambda) \neq 0\}$: the coefficients estimated to be small in the first stage will thus also be encouraged to be small in the second stage, whereas the largest ones will be less penalized.

The variational form of the Lasso can be interpreted as stemming from a hierarchical model in the Bayesian framework (Grandvalet and Canu 1999). In this interpretation, together with λ and the noise variance, the τ_j parameters of Problem (2) define the diagonal covariance matrix of a centered Gaussian prior on $\boldsymbol{\beta}$ (assuming a Gaussian noise model on $\mathbf{y}$). Hence, "freezing" the τ_j parameters at the first stage of a two-stage approach can be interpreted as picking the parameters of the Gaussian prior on $\boldsymbol{\beta}$ to be used at the second stage.

2 A two-stage inference procedure: screen and clean

When interpretability is a key issue, it is essential to take into account the uncertainty associated to the selection of variables inferred from limited data. Indeed, this assessment is critical before investigating possible effects, since there is no way to ascertain that the support is identifiable. Indeed, in practice, the irrepresentable condition and related conditions cannot be checked (Bühlmann 2013).

A classical way to assess the predictor uncertainty consists in testing the significance of each predictor by statistical hypothesis testing and the derived p-values. Although p-values have a number of disadvantages and are prone to possible misinterpretations, it is the numerical indicator that most end-users rely upon when selecting predictors in high-dimensional context. In multiple testing, the most common measures of type-I error are the family wise error rate (FWER) and the false discovery rate (FDR). The FWER is the probability of having at least one false discovery and the FDR is the expected proportion of false discovery among all discoveries. Both criteria, which require reliable p-values as input, are classical alternative, but in applications where numerous tests are performed and where a fairly large proportion of null hypotheses are expected to be false, one is usually prepared to tolerate some type-I errors. Testing with FWER is thus usually considered unduly conservative in biomedical and genomic research, and FDR, which tolerates a proportion of false positives, is appealing in this context (Dudoit and Van der Laan 2008).

Well-established and routinely used selection methods in genomics are univariate (Balding 2006). Although more powerful, multivariate approaches suffer from instability and lack of usual measure of uncertainty. The attempts to assess uncertainty of the Lasso coefficients follow different paths. A first greedy method consists in running permutation tests, mimicking the null hypothesis that the data set is non-informative. This approach may prove computationally heavy and is not trivial to justify from a theoretical point of view (Chatterjee and Lahiri 2013). Bayesian approaches (Kyung et al. 2010) provide an alternative by means of credible intervals for each coefficient. Zhang and Zhang (2014) define a low-dimensional projection estimator, following the efficient score function approach from semi-parametric statistics. Lockhart et al. (2014) propose a test statistic based on Lasso fitted values. This so-called covariance statistic relies on the estimation of the noise variance, whose estimation is problematic for high-dimensional data. Here, we build on Wasserman and Roeder (2009), whose procedure, detailed below, was later extended by Meinshausen et al. (2009) using resampling and an aggregation of p-values for the FWER control. We propose to introduce adaptive ridge in the cleaning stage to transfer more information from the screening stage to the cleaning stage, and thus to make a more extensive use of the subsample of the original data reserved for screening purposes. From a practical point of view, these developments are essential for convincing practitioners of the benefits of multivariate sparse regression models (Boulesteix and Schmid 2014).

2.1 Original screen and clean procedure

The procedure considers a series of sparse models $\{\mathscr{F}_\lambda\}_{\lambda \in \Lambda}$, indexed by a parameter $\lambda \in \Lambda$, which may represent a penalty parameter for regularization methods or a size constraint for subset selection methods. The screening stage consists of two steps. In the first step, each model $\mathscr{F}_\lambda$ is fitted to (part of) the data, thereby selecting a set of possibly relevant variables, that is, the support of the model $\mathscr{S}_\lambda$. Then, in the second step, a model selection procedure chooses a single model $\mathscr{F}_{\hat{\lambda}}$ with its associated $\mathscr{S}_{\hat{\lambda}}$. Next, the cleaning stage eliminates possibly irrelevant variables from $\mathscr{S}_{\hat{\lambda}}$, resulting in the set $\hat{\mathscr{S}}$ that provably controls the type one error rate. The original procedure relies on three independent subsamples of the original data $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2 \cup \mathscr{D}_3$, so as to ensure the consistency of the overall process. The following chart summarizes this procedure, showing the actual use of data that is made at each step:

$$\overbrace{\{1, \ldots, p\} \xrightarrow[(\mathscr{D}_1)]{\substack{\text{step I} \\ \text{fit model}}} \{\mathscr{S}_\lambda\}_{\lambda \in \Lambda} \xrightarrow[(\mathscr{D}_1, \mathscr{D}_2)]{\substack{\text{step II} \\ \text{select model}}}}^{\text{screening stage}} \mathscr{S}_{\hat{\lambda}} \overbrace{\xrightarrow[(\mathscr{D}_3)]{\substack{\text{step III} \\ \text{test support}}} \hat{\mathscr{S}}}^{\text{cleaning stage}} .$$

Under suitable conditions, the screen and clean procedure performs consistent variable selection, that is, it asymptotically recovers the true support with probability one. The two main assumptions are that the screening stage should asymptotically avoid false negatives, and that the size of the true support should be constant, while the number of candidate variables is allowed to grow logarithmically in the number of examples. These assumptions are respectively described in more rigorous terms as the "screening property" and "sparsity property" by Meinshausen et al. (2009).

Empirically, Wasserman and Roeder (2009) tested the procedure with the Lasso, univariate testing, and forward stepwise regression at step I of the screening stage. At step II, model selection was based on ordinary least squares (OLS) regression. The OLS parameters were adjusted on the "training" subsample $\mathscr{D}_1$, using the variables in $\{\mathscr{S}_\lambda\}_{\lambda \in \Lambda}$, and model selection consisted in minimizing the empirical error on the "validation" subsample $\mathscr{D}_2$ with respect to λ. Cleaning was then finally performed by testing the nullity of the OLS coefficients using the independent "test" subsample $\mathscr{D}_3$. Wasserman and Roeder (2009) conclude that the variants using multivariate regression (Lasso and forward stepwise) have similar performances, way above univariate testing.

We now introduce the improvements that we propose here at each stage of the process. Our methodological contribution lies at the cleaning stage, but we also introduced minor modifications at the screening stage that have considerable practical outcomes.

2.2 Adaptive-ridge cleaning stage

The original cleaning stage of Wasserman and Roeder (2009) is based on the ordinary least square (OLS) estimate. This choice is amenable to efficient exact testing procedure for selecting the relevant variables, where the false discovery rate can be provably controlled. However, this advantage comes at a high price:

– first, the procedure can only be used if the OLS is applicable, which requires that the number of variables $\left|\mathscr{S}_{\hat{\lambda}}\right|$ that passed the screening stage is smaller than the number of examples $|\mathscr{D}_3|$ reserved for the cleaning stage;
– second, the only information retained from the screening stage is the support $\mathscr{S}_{\hat{\lambda}}$ itself. There are no other statistics about the estimated regression coefficients that are transferred to this stage.

We propose to make a more effective use of the data reserved for the screening stage by following the approach described in Sect. 1.2: the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ obtained at the screening stage is transferred to the cleaning stage via the adaptive-ridge penalty term. Adaptive refers here to the adaptation of the penalty term to the data at hand. The penalty metric is adjusted to the "training" subsample $\mathscr{D}_1$, its strength is set thanks to the "validation" subsample $\mathscr{D}_2$, and cleaning is eventually performed by testing the nullity of the adaptive-ridge coefficients using the independent "test" subsample $\mathscr{D}_3$.

The statistics computed from our penalized cleaning stage improve the power of the procedure: we observe a dramatic increase in sensitivity (that is, in true positives) at any false discovery rate (see Fig. 3 of the numerical experiment section). With this improved accuracy also comes more precision: the penalization at the cleaning stage brings the additional benefit of stabilizing the selection procedure, with less variability in sensitivity and false discovery rate. Furthermore, our procedure allows for a cleaning stage remaining in the high-dimensional setup (that is, $\left|\mathscr{S}_{\hat{\lambda}}\right| \gg |\mathscr{D}_3|$).

However, using penalized estimators raises a difficulty for the calibration of the statistical tests derived from these statistics. We resolved this issue through the use of permutation tests.

2.3 Testing the significance of the adaptive-ridge coefficients

Student's t-test and Fisher's F-test are two standard ways of testing the nullity of the OLS coefficients. However, these tests do not apply to ridge regression, for which no exact procedure exists.

Halawa and El Bassiouni (1999) proposed a non-exact t-test, but it can be severely off when the explanatory variables are strongly correlated. For example, Cule et al. (2011) report a false positive rate as high as 32 % for a significance level supposedly fixed at 5 %. Typically, the inaccuracy aggravates with high penalty parameters, due to the bias of the ridge regression estimate, and due to the dependency between the response variable and the ridge regression residuals.

The F-test compares the goodness-of-fit of two nested models. Let $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_0$ be the n-dimensional vectors of predictions for the larger and smaller model respectively. The F-statistic

$$F = \frac{\left\|\mathbf{y} - \hat{\mathbf{y}}_0\right\|^2 - \left\|\mathbf{y} - \hat{\mathbf{y}}_1\right\|^2}{\left\|\mathbf{y} - \hat{\mathbf{y}}_1\right\|^2}, \tag{4}$$

follows a Fisher distribution when $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_0$ are estimated by ordinary least squares under the null hypothesis that the smaller model is correct. Although it is widely used for model selection in penalized regression problems (for calibration and degrees of freedom issues, see, Hastie and Tibshirani 1990), the F-test is not exact for ridge regression, for the reasons already mentioned above—estimation bias

Table 1 Average false positive rate FPR (or type-I error) and sensitivity SEN (or power) computed on 500 simulations over the set of variables passing the screening stage

Design	IND		BLOCK		GROUP		TOEP$^-$	
	FPR	SEN	FPR	SEN	FPR	SEN	FPR	SEN
permutation F-test	5.1	92.4	3.9	86.7	3.9	62.3	4.7	81.9
standard F-test	9.9	93.1	11.8	89.6	14.8	73.0	15.4	87.1
standard t-test	8.0	94.0	12.4	93.1	5.8	95.7	7.9	85.1

The prescribed significance level is 5 %. The IND, BLOCK, GROUP and TOEP$^-$ designs are fully described in Sect. 4.1

and dependency between the numerator and the denominator in Eq. (4). Here, we propose to approach the distribution of the F-statistic under the null hypothesis by randomization. We permute the values taken by the explicative variable to be tested, on the larger model, so as to estimate the distribution of the F-statistic under the null hypothesis that the variable is irrelevant. This permutation test is asymptotically exact when the tested variable is independent from the other explicative variables, and is approximate in the general case.[3] It can be efficiently implemented using block-wise decompositions, thereby saving a factor p, as detailed in Appendix 2.

Table 1 shows that, compared to the standard t-test and F-test (see e.g. Hastie and Tibshirani 1990), the permutation test provides a satisfactory control of the significance level. It is either well-calibrated or slightly more conservative than the prescribed significance level, whereas the standard t-test and F-test result in false positive rates that are way above the asserted significance level, especially for strong correlations between explanatory variables. These observations apply throughout the experiments reported in Section 4.1.

Testing all variables results in a multiple testing problem. We propose here to control the false discovery rate (FDR), which is defined as the expected proportion of false discoveries among all discoveries. This control requires to correct the p-values for multiple testing (Benjamini and Hochberg 1995). The overall procedure is well calibrated as shown in Sect. 4.

2.4 Modifications at screening stage

Wasserman and Roeder (2009) propose to use two subsamples at the cleaning stage in order to establish the consistency of the screen and clean procedure. Indeed, this consistency relies partly on the fact that all relevant variables pass the screening stage with very high probability. This "screening property" (Meinshausen et al. 2009) was established using the protocol described in Sect. 2.1. To our knowledge, it remains to be proved for model selection based on cross-validation. However, Wasserman and Roeder (2009) mention another procedure relying on two independent subsamples of the original data $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2$, where model selection relies on leave-one-out cross-validation on $\mathscr{D}_1$ and $\mathscr{D}_2$ is reserved for cleaning. The following chart summarizes this modified procedure:

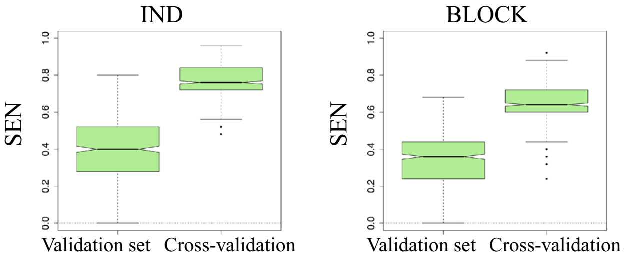

Fig. 1 Sensitivity of the screen and clean procedure (the higher, the better), for the two model selection strategies at the screening stage, and FDR controlled at 5 % based on the permutation test. Lasso regression is used in the screening stage and adaptive-ridge regression in the cleaning stage. Each boxplot is computed based over 500 replications for the IND and BLOCK simulation designs (see Sect. 4.1 for full description)

$$\overbrace{\{1,\ldots,p\}\xrightarrow[(\mathscr{D}_1)]{\substack{\text{step I}\\ \text{fit model}}}\{\mathscr{S}_\lambda\}_{\lambda\in\Lambda}\xrightarrow[(\mathscr{D}_1)]{\substack{\text{step II}\\ \text{select model}}}\mathscr{S}_{\hat{\lambda}}}^{\text{screening stage}}\ \overbrace{\xrightarrow[(\mathscr{D}_2)]{\substack{\text{step III}\\ \text{test support}}}\hat{\mathscr{S}}}^{\text{cleaning stage}}\ .$$

Hence, half of the data are now devoted to each stage of the method. We followed here this variant, which results in important sensitivity gains for the overall selection procedure, as illustrated in Fig. 1.

We slightly depart from (Wasserman and Roeder 2009), by selecting the model by 10-fold cross-validation, and, more importantly, by using the sum of squares residuals of the *penalized* estimator for model selection. Note that Wasserman and Roeder (2009), and later Meinshausen et al. (2009) based model selection on the OLS estimate using the support $\mathscr{S}_\lambda$. This choice implicitly limits the size of the selected support $|\mathscr{S}_{\hat{\lambda}}| < \frac{n}{2}$, which is actually implemented more stringently as $|\mathscr{S}_{\hat{\lambda}}| \leq \sqrt{n}$ and $|\mathscr{S}_{\hat{\lambda}}| \leq \frac{n}{6}$ by Wasserman and Roeder (2009) and Meinshausen et al. (2009) respectively. Our model selection criterion allows for more variables to be transferred to the cleaning stage, so that the screening property is more likely to hold.

[3] Note that there is no finite-sample exact permutation test in multiple linear regression (Anderson and Robinson 2001). A test based on partial residuals (under the null hypothesis regression model) is asymptotically exact for unpenalized regression, but it does not apply to penalized regression.

3 Analysis of the orthonormal design

Here, we propose a detailed analysis of the benefits of the procedure in the orthonormal design, where we assume that we have two samples $\mathscr{D}_1$ and $\mathscr{D}_2$ of size n with design matrices $n^{-1}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$. In this situation, the screening stage based on the Lasso provides

$$\hat{\beta}_j^{\text{screen}}(\lambda) = \left(1 - \frac{\lambda}{n|\hat{\beta}_j^{\text{ols}}|}\right)_+ \hat{\beta}_j^{\text{ols}},$$

where $\hat{\boldsymbol{\beta}}^{\text{ols}}$ is the ordinary least squares estimator (Tibshirani 1996). Assuming additionally a Gaussian noise in the model, $\boldsymbol{\varepsilon} \sim \mathscr{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, the probability that variable j does not pass the screening stage is:

$$\begin{aligned} P(\beta_j^\star, \lambda) &= \mathbb{P}\left[\hat{\beta}_j^{\text{screen}}(\lambda) = 0\right] \\ &= \Phi\left(\frac{n^{1/2}}{\sigma}\left(\frac{\lambda}{n} - \beta_j^\star\right)\right) - \Phi\left(-\frac{n^{1/2}}{\sigma}\left(\frac{\lambda}{n} + \beta_j^\star\right)\right), \end{aligned}$$

where Φ is the cumulative distribution of the standard normal distribution. Then, as the cleaning stage operates on an independent sample, the distributions for the cleaning stage estimators are:

$$f(\hat{\beta}_j^{\text{clean}}) = P(\beta_j^\star, \lambda)\, \delta(\hat{\beta}_j^{\text{clean}}) + (1 - P(\beta_j^\star, \lambda))\, g(\hat{\beta}_j^{\text{clean}}),$$

with

$$g(\hat{\beta}_j^{\text{clean}}) = \frac{n^{1/2}}{\sigma} \varphi\left(\frac{n^{1/2}}{\sigma}(\hat{\beta}_j^{\text{clean}} - \beta_j^\star)\right)$$

for the OLS estimator, and

$$\begin{aligned} g(\hat{\beta}_j^{\text{clean}}) = \frac{n^{1/2}}{\sigma} \int_0^1 x^{-1} \varphi \\ \times \left(x^{-1} \frac{n^{1/2}}{\sigma}(\hat{\beta}_j^{\text{clean}} - \beta_j^\star)\right) h_j(x)\, dx \end{aligned}$$

for the adaptive-ridge estimator (AR), where δ is the Dirac delta function, φ is the standard normal distribution, and h_j is the distribution of the shrinkage coefficient that is applied to variable j by AR, that is, the distribution of $|\hat{\beta}_j^{\text{screen}}|/(|\hat{\beta}_j^{\text{screen}}| + n^{-1}\lambda)$. There is no simple analytical form of the overall distribution for the adaptive-ridge estimator, but these formulas are however interesting, in that they exhibit the three parameters of importance for the distribution of the cleaning estimators, namely $\beta_j^\star$, $n^{1/2}\sigma^{-1}$, and $n^{-1}\lambda$. Furthermore, in all expressions, $n^{1/2}\sigma^{-1}$ acts as a scale parameter. We can therefore provide a scale-free analysis by studying the role of the normalized penalty parameter $n^{-1/2}\sigma^{-1}\lambda$ on the normalized estimator $n^{1/2}\sigma^{-1}\hat{\beta}_j^{\text{clean}}$, when the normalized true parameter $n^{1/2}\sigma^{-1}\beta_j^\star$ varies.

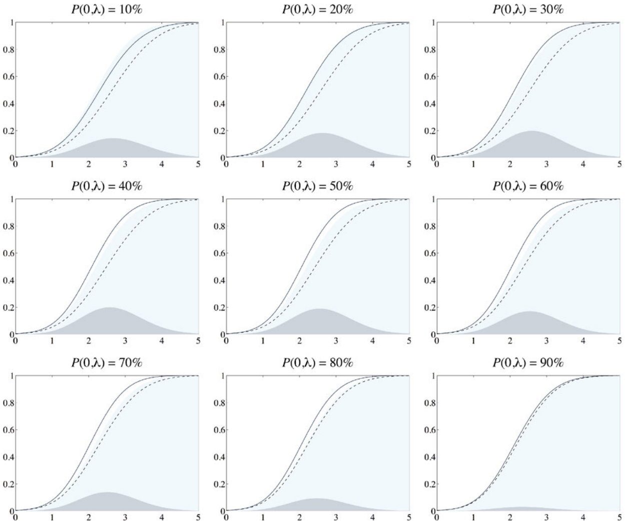

Fig. 2 Power (or sensitivity) as a function of $n^{1/2}\sigma^{-1}\beta_j^\star$, for a univariate test at the 1 % level based on OLS cleaning (*dashed*) or AR cleaning (*plain*). The *light gray area* in the bottom displays the difference between the two curves, and the boundary of the very *light blue area*, included for cross-reference, represents the best result achieved using the OLS cleaning estimator, for $P(0, \lambda) = 90$ %. (Color figure online)

We now make use of these observations to compare the power of the statistical testing of the nullity of $\beta_j^\star$ from the OLS and from the AR cleaning estimator. First, the expected type-I-error is fixed to 1 % using the distributions of $\hat{\beta}_j^{\text{clean}}$ for the OLS and the AR estimator for $\beta_j^\star = 0$, and we then compute the expected type-II-error according to $n^{1/2}\sigma^{-1}\beta_j^\star$. A significance level of 1 % roughly corresponds to the effective significance level for unitary tests in our experiments of Sect. 4.2 aiming at controlling the FDR at 5 % using the Benjamini–Hochberg procedure.

Figure 2 represents graphs spanning the possible values of $n^{-1/2}\sigma^{-1}\lambda$. For readability, we indexed subfigures by $P(0, \lambda)$, the probability that a null variable is filtered at screening stage. We observe that, for any λ value, the test based on the AR estimator has uniformly higher power than the one based on the OLS estimator. Furthermore, for most λ values, AR cleaning performs better than the best λ setting for OLS cleaning. This means that AR cleaning often brings more than having an oracle for selecting λ at the screening stage.

4 Numerical experiments

Variable selection algorithms are difficult to assess objectively on real data, where the truly relevant variables are unknown. Simulated data provide a direct access to the ground truth, in a situation where the statistical hypotheses hold. In this section, we first analyze the performances of our variable selection method on simulations, before presenting an application to a Genome Wide Association case Study on HIV-1 infection.

4.1 Simulation models

We consider the linear regression model $Y = X\boldsymbol{\beta}^\star + \varepsilon$, where Y is a continuous response variable, $X = (X_1, \ldots, X_p)$ is a vector of p predictor variables, $\boldsymbol{\beta}^\star$ is the vector of unknown parameters and ε is a zero-mean Gaussian error variable with variance σ^2. The parameter $\boldsymbol{\beta}^\star$ is sparse, that is, the support set $\mathscr{S}^\star = \left\{ j \in \{1, \ldots, p\} | \beta_j^\star \neq 0 \right\}$ indexing its non-zero coefficients is small $|\mathscr{S}^\star| \ll p$.

Variable selection is known to be affected by numerous factors: the number of examples n, the number of variables p, the sparseness of the model $|\mathscr{S}^\star|$, the correlation structure of the explicative variables, the relative magnitude of the relevant parameters $\{\beta_j^\star\}_{j \in \mathscr{S}^\star}$, and the signal-to-noise ratio SNR.

In our experiments, we varied $n \in \{250, 500\}$, $p \in \{250, 500\}$, $|\mathscr{S}^\star| \in \{25, 50\}$, $\rho \in \{0.5, 0.8\}$. We considered four predictor correlation structures:

IND independent explicative variables following a zero-mean, unit-variance Gaussian distribution: $X \sim \mathscr{N}(\mathbf{0}, \mathbf{I}_p)$;

BLOCK dependent explicative variables following a zero-mean Gaussian distribution, with a block-diagonal covariance matrix: $X \sim \mathscr{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for all pairs (i, j), $j \neq i$ belonging to the same block and $\Sigma_{ij} = 0$ for all pairs (i, j) belonging to different blocks. Each block comprises 25 variables.
The position of relevant variables is dissociated from the block structure, that is, randomly distributed in $\{1, \ldots, p\}$. This design is thus difficult for variable selection.

GROUP same as BLOCK, except that the relevant variables are gathered a single block when $|\mathscr{S}^\star| = 25$ and in two blocks when $|\mathscr{S}^\star| = 50$, thus facilitating group variable selection.

TOEP$^-$ same as GROUP, except that $\Sigma_{ij} = -\rho^{|i-j|}$ for all pairs (i, j), $j \neq i$ belonging to the same block and $\Sigma_{ij} = 0$ for all pairs (i, j) belonging to different blocks.
This design allows for strong negative correlations.

The non-zero parameters $\beta_j^\star$ are drawn from a uniform distribution $\mathscr{U}(10^{-1}, 1)$ to enable different magnitudes. Finally, the signal to noise ratio, defined as $\text{SNR} = \boldsymbol{\beta}^{\star\top}\boldsymbol{\Sigma}\boldsymbol{\beta}^\star/\sigma^2$ varies in $\{4, 8, 32\}$.

4.2 Results

In the following, we discuss the IND, BLOCK, GROUP and TOEP$^-$ designs with $n = 250$, $p = 500$, $|\mathscr{S}^\star| = 25$, $\rho = 0.5$ and $\text{SNR} = 4$, since the relative behavior of all methods is representative of the general pattern that we observed across all simulation settings. These setups lead to feasible but challenging problems for model selection.

All variants of the screen and clean procedure are evaluated here with respect to their sensitivity (SEN), for a controlled false discovery rate FDR. These two measures are the analogs of power and significance in the single hypothesis testing framework:

$$\text{SEN} = \mathbb{E}\left[\frac{TP}{TP + FN}\mathbb{I}_{\{(TP+FN)>0\}}\right]$$
$$\text{FDR} = \mathbb{E}\left[\frac{FP}{TP + FP}\mathbb{I}_{\{(TP+FP)>0\}}\right],$$

where FP is the number of false positives, TP is the number of true positives and FN is the number of false negatives.

We first show the importance of the cleaning stage for FDR control. We then show the benefits of our proposal compared to the original procedure of Wasserman and Roeder (2009) and to the univariate approach. The variable selection method of Lockhart et al. (2014) was not included in these experiments, because it did not produce convincing results in these small n large p designs where the noise variance is not assumed to be known.

Importance of the cleaning stage Table 2 shows that the cleaning step is essential to control the FDR at the desired level. In the screening stage, the variables selected by the Lasso are way too numerous: first, the penalty parameter is determined to optimize the cross-validated mean squared

Table 2 False discovery rate FDR and sensitivity SEN, computed over 500 simulations for each design. The screening stage (w/o cleaning) is not calibrated; the cleaning stage is calibrated to control the FDR below 5 %, using the Benjamini–Hochberg procedure

Design	IND		BLOCK		GROUP		TOEP$^-$	
	FDR	SEN	FDR	SEN	FDR	SEN	FDR	SEN
W/o cleaning	76.7	87.5	76.0	83.9	38.9	86.2	79.9	56.5
AR cleaning	4.2	76.1	2.8	64.8	1.7	37.7	4.3	39.6
OLS cleaning	3.9	48.3	3.1	37.1	2.5	17.9	3.7	25.3
Univar	4.4	40.4	86.4	71.0	5.3	100.0	4.2	28.4

Our adaptive-ridge (AR) cleaning is compared with the original (OLS) cleaning and univariate testing (Univar)

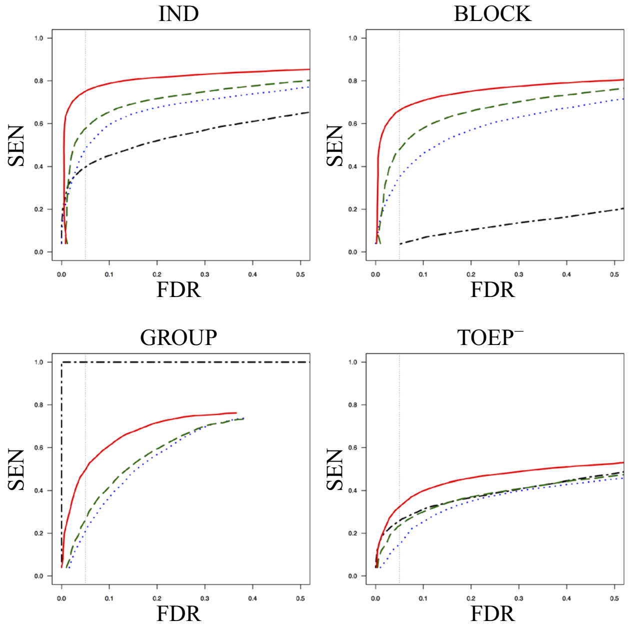

Fig. 3 Sensitivity SEN versus False Discovery Rate FDR (the higher, the better). Lasso screening followed by: adaptive-ridge cleaning (*red solid line*), ridge cleaning (*green dashed line*), OLS cleaning (*blue dotted line*); univariate testing (*black dot-dashed line*). All curves are indexed by the rank of the test statistics, and averaged over the 500 simulations of each design. The *vertical dotted line* marks the 5 % FDR level. (Color figure online)

error, which is not optimal for model selection; second, we are far from the asymptotic regime where support recovery can be achieved. As a result, the Lasso performs rather poorly. Cleaning enables the control of the FDR, leading of course to a decrease in sensitivity, which is moderate for independent variables, and higher in the presence of correlations.

Comparisons of controlled selection procedures Figure 3 provides a global picture of sensitivity according to FDR, for the test statistics computed in the cleaning stage. First, we observe that the direct univariate approach, which simply considers a t-statistic for each variable independently, is by far the worst option in the IND, BLOCK and TOEP^{-} designs, and by far the best in the GROUP design. In this last situation, the univariate approach confidently detects all the correlated variables of the relevant group, while the regression-based approaches are hindered by the high level of correlation between variables. Betting on the univariate approach may thus be profitable, but it is also risky due to its extremely erratic behavior.

Second, we see that our adaptive-ridge cleaning systematically dominates the original OLS cleaning. In this respect, experiments meet the analysis of Sect. 3, but our experimental results are even more strongly in favor of adaptive-ridge cleaning. There is thus an important practical additional benefit of adaptive-ridge cleaning which cannot be explained solely by the analysis of Sect. 3. To isolate the effect of transfering the magnitude of weights from the effect of the regularization brought by adaptive-ridge, we computed the results obtained from a cleaning step based on plain ridge regression (with regularization parameter set by cross-validation). We see that ridge regression cleaning improves upon OLS cleaning, and that adaptive-ridge cleaning brings this improvement much further, thus confirming the practical value of the weight transfer from the screening stage to the cleaning stage. Note that in the orthonormal design of Sect. 3, ridge regression behaves exactly as OLS regarding the power curves of tests. Indeed, since all parameters are equally shrunk in the orthonormal setting, $\hat{\boldsymbol{\beta}}^{\text{clean}}$ for OLS and ridge only differ by a shrinking constant, which does not impact the performances of the tests.

Table 2 shows the actual operating conditions of the variable selection procedures, when a threshold on the test statistics has to be set to control the FDR. Here, the threshold is set to control the FDR at the 5 % level, using the Benjamini–Hochberg procedure. This control is always effective for the screen and clean procedures, but not for variable selection based on univariate testing. In all designs, our proposal dramatically improves over the original OLS strategy, with sensitivity gains ranging from 50 to 100 %. All differences in sensitivity are statistically significant. The variability of FDP and sensitivity is not shown to avoid clutter, but the smallest variability in FDP is obtained for the adaptive-ridge cleaning, while the smallest variability in sensitivity is obtained for univariate regression, followed by adaptive-ridge cleaning. The adaptive ridge penalty thus brings about more stability to the overall selection process.

4.3 GWAS on HIV

We now compare the results of variable selection in a Genome Wide Association Study (GWAS) on HIV-1 infection (Dalmasso et al. 2008). One of the goal of this study was to identify genomic regions that influence HIV-RNA levels during primary infection. Genotypes from $n = 605$ seroconverters were obtained using Illumina Sentrix Human Hap300 Beadchips. As different subregions of the major histocompatibility complex (MHC) had been shown to be associated with HIV-1 disease, the focus is set on the $p = 20,811$ Single Nucleotide Polymorphisms (SNPs) located on Chromosome 6. The $20,811$ explanatory variables are categorical variables with three levels, encoded as 1 for homozygous samples "AA", 2 for heterozygous samples "AB" and 3 for homozygous samples "BB" (where "A" and "B" correspond to the two possible alleles for each SNP). The quantitative response variable is the plasma HIV-RNA level, which is a marker of the HIV disease progression.

Table 3 Adjusted p-values (in %) obtained from the Benjamini–Hochberg procedure for the five SNPs of the HIV data selected at a 25 % FDR level

SNP	Genomic Region	AR cleaning	OLS cleaning	Univar
rs2523619	MHC	0.0	2.3	0.2
rs214590	MHC	21.0	38.9	12.7
rs11967684	MHC	21.0	68.5	10.3
rs6923486	other	21.0	42.6	99.5
rs1983789	other	24.8	39.9	96.2

Our adaptive-ridge (AR) cleaning is compared with the original (OLS) cleaning and with univariate testing (Univar)

The Lasso screening selects $|\mathscr{S}_{\hat{\lambda}}| = 20$ SNPs. Considering a 25 % FDR level (as in, Dalmasso et al. 2008), the adaptive-ridge screening selects $|\hat{\mathscr{S}}| = 5$ SNPs as being associated with the plasma HIV-RNA, while OLS selects only $|\hat{\mathscr{S}}| = 1$ of them (see Table 3). Among the 12 SNPs which were identified by Dalmasso et al. (2008) from a univariate analysis in the MHC region, only 3 (rs2523619, rs214590 and rs11967684) remain selected with the proposed approach, and only one with the OLS cleaning. It is worth noting that these 12 SNPs can be clustered into two groups with high positive intra-block correlations and high negative inter-block correlations (up to $|\rho| = 0.7$). Hence, there is a high chance of confusion between these highly correlated variables. In this situation, variable selection methods working on sets of variables, such as the ones we envision in future works would be highly valuable. Those results are in line with the simulation study, in the sense that, in a similar context, the adaptive-ridge cleaning stage has a better sensitivity than OLS cleaning and is also much more conservative than univariate testing.

5 Discussion

We propose to use the magnitude of regression coefficients in two-stage variable selection procedures. We use the connection between the Lasso and adaptive-ridge (Grandvalet 1998) to convey more information from the first stage to the second stage: the magnitude of the coefficients estimated at the first stage is transferred to the second stage through penalty parameters. Our proposal results in a new "screen and clean" procedure (Wasserman and Roeder 2009) for assessing the uncertainties pertaining to the selection of relevant variables in regression problems.

On the theoretical side, our analysis in the orthonormal setting (which is numerical but precise and accurate up to numerical integration errors) shows that our cleaning stage produces test statistics that systematically dominate the ones of the original cleaning stage based on OLS regression. The benefits are comparable with the ones of having an oracle for the penalty parameter in the first stage of the procedure.

Empirically, our procedure controls the False Discovery Rate, even in difficult settings, with high correlations between variables. Furthermore, the sensitivity obtained by our cleaning stage is always as good, and often much better than the one based on the ordinary least squares. Part of this improvement is due to the stabilization effect of the penalization, but the benefit of the adaptive penalty shown in the orthonormal setting is observed in practice in all settings. The penalized cleaning step also allows for a less severe screening, since cleaning can then handle more than $n/2$ variables. Our procedure can thus be employed in very high-dimensional settings, as the screening property (that is, in the words of Bühlmann (2013), the ability of the Lasso to select all relevant variables) is more easily fulfilled, which is essential for a reliable control of the false discovery rate.

Several interesting directions are left for future works. The connection between the two stages can be generalized to all penalties for which a quadratic variational formulation can be derived. This is particularly appealing for structured penalties such as the fused-lasso or the group-Lasso, allowing to use the knowledge of groups at the second stage, through penalization coefficients.

Acknowledgements This work was supported by the UTC foundation for Innovation, in the ToxOnChip program. It has been carried out in the framework of the Labex MS2T (ANR-11-IDEX-0004-02) within the "Investments for the future" program, managed by the National Agency for Research.

Variational equivalence

We show below that the quadratic penalty in $\boldsymbol{\beta}$ in Problem (2) acts as the Lasso penalty $\lambda \|\boldsymbol{\beta}\|_1$.

Proof The Lagrangian of Problem (2) is:

$$L(\boldsymbol{\beta}) = J(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \frac{1}{\tau_j} \beta_j^2 + \nu_0 \left(\sum_{j=1}^{p} \tau_j - \|\boldsymbol{\beta}\|_1 \right) - \sum_{j=1}^{p} \nu_j \tau_j .$$

Thus, the first order optimality conditions for τ_j are

$$\begin{aligned} \frac{\partial L}{\partial \tau_j}(\tau_j^\star) = 0 &\Leftrightarrow -\lambda \frac{\beta_j^2}{\tau_j^{\star 2}} + \nu_0 - \nu_j = 0 \\ &\Leftrightarrow -\lambda \beta_j^2 + \nu_0\, \tau_j^{\star 2} - \nu_j\, \tau_j^{\star 2} = 0 \\ &\Rightarrow -\lambda \beta_j^2 + \nu_0\, \tau_j^{\star 2} = 0, \end{aligned}$$

where the term in ν_j vanishes due to complementary slackness, which implies here $\nu_j \tau_j^\star = 0$. Together with the constraints of Problem (2), the last equation implies $\tau_j^\star = |\beta_j|$, hence Problem (2) is equivalent to

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 ,$$

which is the original Lasso formulation. □

Efficient implementation

Permutation tests rely on the simulation of numerous data sampled under the null hypothesis distribution. The number of replications must be important to estimate the rather extreme quantiles we are typically interested in. Here, we use $B = 1000$ replications for the $q = |\mathscr{S}_{\hat{\lambda}}|$ variables selected in the screening stage. Each replication involving the fitting of a model, the total computational cost for solving these B systems of size q on the q selected variables is $O(Bq(q^3 + q^2 n))$. However, block-wise decompositions and inversions can bring computing savings by a factor q.

First, we recall that the adaptive-ridge estimate, computed at the cleaning stage, is computed as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\boldsymbol{\Lambda}$ is the diagonal adaptive-penalty matrix defined at the screening stage, whose jth diagonal entry is $\lambda/\tau_j^\star$, as defined in (1–3).

In the F-statistic (4), the permutation affects the calculation of the larger model $\hat{\mathbf{y}}_1$, which is denoted $\hat{\mathbf{y}}_1^{(b)}$ for the bth permutation. Using a similar notation convention for the design matrix and the estimated parameters, we have $\hat{\mathbf{y}}_1^{(b)} = \mathbf{X}^{(b)}\hat{\boldsymbol{\beta}}^{(b)}$. When testing the relevance of variable j, $\mathbf{X}^{(b)}$ is defined as the concatenation of the permuted variable $\mathbf{x}_j^{(b)}$ and the other original variables: $\mathbf{X}^{(b)} = (\mathbf{x}_j^{(b)}, \mathbf{X}_{\setminus j}) = (\mathbf{x}_j^{(b)}, \mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ..., \mathbf{x}_p)$. Then, $\hat{\boldsymbol{\beta}}^{(b)}$ can be efficiently computed by using $a^{(b)} \in \mathbb{R}$, $\mathbf{v}^{(b)} \in \mathbb{R}^{q-1}$ and $\hat{\boldsymbol{\beta}}_{\setminus j} \in \mathbb{R}^{q-1}$ defined as follows:

$$\begin{aligned}
a^{(b)} &= (\|\mathbf{x}_j^{(b)}\|_2^2 + \Lambda_{jj}) - \mathbf{x}_j^{(b)\top}\mathbf{X}_{\setminus j}(\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1}\mathbf{X}_{\setminus j}^\top\mathbf{x}_j^{(b)}\\
\mathbf{v}^{(b)} &= -(\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1}\mathbf{X}_{\setminus j}^\top\mathbf{x}_j^{(b)}\\
\hat{\boldsymbol{\beta}}_{\setminus j} &= (\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1}\mathbf{X}_{\setminus j}^\top\mathbf{y}.
\end{aligned}$$

Indeed, using the Schur complement, one writes $\hat{\boldsymbol{\beta}}^{(b)}$ as follows:

$$\hat{\boldsymbol{\beta}}^{(b)} = \frac{1}{a^{(b)}}\begin{pmatrix}1\\ \mathbf{v}^{(b)}\end{pmatrix}\begin{pmatrix}1 & \mathbf{v}^{(b)\top}\end{pmatrix}\begin{pmatrix}\mathbf{x}_j^{(b)\top}\mathbf{y}\\ \mathbf{X}_{\setminus j}^\top\mathbf{y}\end{pmatrix} + \begin{pmatrix}0\\ \hat{\boldsymbol{\beta}}_{\setminus j}\end{pmatrix}.$$

Hence, $\hat{\boldsymbol{\beta}}^{(b)}$ can be obtained as a correction of the vector of coefficients $\hat{\boldsymbol{\beta}}_{\setminus j}$ obtained under the smaller model. The key observation to be made here is that $\mathbf{x}_j^{(b)}$ does not intervene in the expression $(\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1}$, which is the bottleneck in the computation of $a^{(b)}$, $\mathbf{v}^{(b)}$ and $\hat{\boldsymbol{\beta}}_{\setminus j}$. It can therefore be performed once for all permutations. Additionally, $(\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1}$ can be cheaply computed from $\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}$ as follows:

$$\begin{aligned}
(\mathbf{X}_{\setminus j}^\top\mathbf{X}_{\setminus j} + \boldsymbol{\Lambda}_{\setminus j})^{-1} &= \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\setminus j \setminus j}\\
&\quad - \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\setminus jj}\left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{jj}^{-1}\\
&\quad \times \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{j\setminus j}.
\end{aligned}$$

Thus we compute $\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}$ once, firstly correct for the removal of variable j, secondly correct for permutation b, thus eventually requiring $O(B(q^3 + q^2 n)))$ operations.

References

Ambroise, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. **99**(10), 6562–6566 (2002)

Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol. **11**(10), R106 (2010)

Anderson, M.J., Robinson, J.: Permutation tests for linear models. Austral. N. Z. J. Stat. **43**(1), 75–88 (2001)

Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Found. Trends Mach. Learn. **4**(1), 1–106 (2012)

Balding, D.: A tutorial on statistical methods for population association studies. Nat. Rev. Genet. **7**(10), 781–791 (2006)

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

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**(1), 289–300 (1995)

Boulesteix, A.L., Schmid, M.: Machine learning versus statistical modeling. Biom. J. **56**, 588–593 (2014)

Bühlmann, P.: Statistical significance in high-dimensional linear models. Bernoulli **19**, 1212–1242 (2013)

Candès, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)

Chatterjee, A., Lahiri, S.N.: Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. Ann. Stat. **41**(3), 1232–1259 (2013)

Chong, I.G., Jun, C.H.: Performance of some variable selection methods when multicollinearity is present. Chemom. Intel. Lab. Syst. **78**(1–2), 103–112 (2005)

Cule, E., Vineis, P., De Lorio, M.: Significance testing in ridge regression for genetic data. BMC Bioinf. **12**(372), 1–15 (2011)

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M.L., Lambotte, O., Avettand-Fenoel, V., Le Clerc, S., Denis de Senneville, L., Deveau, C., Boufassa, F., Debre, P., Delfraissy, J.F., Broet, P., Theodorou, I.: Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS genome wide association 01 study. PLoS One **3**(12), e3907 (2008)

Dudoit, S., Van der Laan, M.: Multiple Testing Procedures with Applications to Genomics. Springer, New York (2008)

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)

Grandvalet, Y.: Least absolute shrinkage is equivalent to quadratic penalization. In: Niklasson L, Bodén M, Ziemske T (eds) ICANN'98, Perspectives in Neural Computing, vol 1, Springer, New York, pp. 201–206 (1998)

Grandvalet, Y., Canu, S.: Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In: Kearns MS, Solla SA, Cohn DA (eds) Advances in Neural Information Processing Systems 11 (NIPS 1998), MIT Press, Cambridge, pp. 445–451 (1999)

Halawa, A.M., El Bassiouni, M.Y.: Tests of regressions coefficients under ridge regression models. J. Stat. Comput. Simul. **65**(1), 341–356 (1999)

Hastie, T.J., Tibshirani, R.J.: Generalized Additive Models, Monographs on Statistics and Applied Probability, vol. 43. Chapman & Hall, London (1990)

Huang, J., Horowitz, J.L., Ma, S.: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Stat. **36**(2), 587–613 (2008)

Kyung, M., Gill, J., Ghosh, M., Casella, G.: Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. **5**(2), 369–411 (2010)

Liu, H., Yu, B.: Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. Electr. J. Stat. **7**, 3124–3169 (2013)

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R.: A significance test for the lasso. Ann. Stat. **42**(2), 413–468(2014)

Meinshausen, N.: Relaxed lasso. Comput. Stat. Data Anal. **52**(1), 374–393 (2007)

Meinshausen, N., Meier, L., Bühlmann, P.: *p*-values for high-dimensional regression. J. Am. Stat. Assoc. **104**(488), 1671–1681 (2009)

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.A., Grill, J., Frouin, V.: Variable selection for generalized canonical correlation analysis. Biostatistics **15**(3), 569–583 (2014)

Tibshirani, R.J.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)

Verzelen, N.: Minimax risks for sparse regressions: ultra-high dimensional phenomenons. Electr. J. Stat. **6**, 38–90 (2012)

Wang, Y., Yang, J., Yin, W., Zhang, W.: A new alternating minimization algorithm for total variation image reconstruction. SIAM J. Imaging Sci. **1**(3), 248–272 (2008)

Wasserman, L., Roeder, K.: High-dimensional variable selection. Ann. Stat. **37**(5A), 2178–2201 (2009)

Xing, E.P., Jordan, M.I., Karp, R.M.: Feature selection for high-dimensional genomic microarray data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), pp. 601–608 (2001)

Zhang, C.H., Zhang, S.S.: Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76**(1), 217–242 (2014)