Significance Testing for Variable Selection in High-Dimension

Jean-Michel Bécu, Christophe Ambroise, Yves Grandvalet, Cyril Dalmasso

***Abstract*—Assessing the uncertainty pertaining to the conclusions derived from experimental data is challenging when there is a high number of possible explanations compared to the number of experiments. We propose a new two-stage "screen and clean" procedure for assessing the uncertainties pertaining to the selection of relevant variables in high-dimensional regression problems. In this two-stage method, screening consists in selecting a subset of candidate variables by a sparsity-inducing penalized regression, while cleaning consists in discarding all variables that do not pass a significance test. This test was originally based on ordinary least squares regression. We propose to improve the procedure by conveying more information from the screening stage to the cleaning stage. Our cleaning stage is based on an adaptively penalized regression whose weights are adjusted in the screening stage. Our procedure is amenable to the computation of p-values, allowing to control the False Discovery Rate. Our experiments show the benefits of our procedure, as we observe a systematic improvement of sensitivity compared to the original procedure.**

***Keywords*—False discovery rate, Genome Wide Association Study, Linear model, Lasso, p-values, Screen and clean, Variable selection**

I. 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 [1], chemometry [2] and genomics [3]–[5].

Since the development of the sparse estimators derived from ℓ_1 penalties such as the Lasso [6] or the Dantzig selector [7], sparse models have been shown to be able to recover the subset of relevant variables in various situations [7]–[10]. That being said, the conditions for support recovery are quite stringent and difficult to assess in practice. Hence, it is essential to take into account the uncertainty associated to the selection of variables inferred from limited data when interpretability is a key issue.

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 biologists or physicians rely upon when selecting predictors in high-dimensional context. Well-established and routinely used selection methods in genomics are univariate [11]. Although more powerful, multivariate approaches suffer from instability and lack of usual measure of uncertainty. It is only recently that means for computing p-values or confidence intervals in the high-dimensional regression setup were proposed, originating with the work of *Wasserman et al.* [12] and followed by others [9], [13], [14].

From a practical point of view, these recent developments are essential to convince practitioners of the benefit of multivariate sparse regression models. This paper contributes to this effort. Section 2 summarizes the state-of-the-art in the computation of significance test for the Lasso in high dimension. In Section 3, we propose an original approach for testing the significance of the Lasso and the Elastic-Net regression coefficients, relying on a variational form of these estimates. The numerical experiments presented in Section 4 demonstrate dramatic gains in performance compared to the original procedure of *Wasserman et al.* [12] or to the classical univariate testing approach.

II. High-Dimensional Variable Selection via Sparse Regression

A. Overview

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

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

where $\mathbf{y} = (y_1, \cdots, y_n)^t$ 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.

Several theoretical results state conditions that ensure asymptotical support recovery, that is, the recovery of the subset of all relevant explanatory variables.One of the main result reveals a necessary and sufficient condition for the selection property of ℓ_1-regularized least squares. Several variants of this condition have been proposed, either known as the irrepresentable condition, or the mutual incoherence condition in signal processing. In a nutshell, this type of condition states that the subset of truly effective variables can be retrieved exactly, provided the relevant and irrelevant covariates are not too strongly correlated. However, in practice, the irrepresentable condition and related conditions cannot be checked [9]. Hence, there is no way to ascertain that the support is identifiable.

J.-M. Bécu and Y. Grandvalet are with Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc UMR 7253, CS 60 319, 60 203 Compiègne cedex, France e-mail: jean-michel.becu@hds.utc.fr.

C. Ambroise and C. Dalmasso are with the LaMME, Université d'Évry val d'Essonne, 91000 Évry, France.

978-1-4799-6926-5/15/$31.00 ©2015 European Union

In the impossibility to ensure that the selected predictors are part of the true support, it seems reasonable to further test the nullity of the regression coefficients. This problem suggests the use of the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR) as the type-I multiple testing error. 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 [15].

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 [16]. Bayesian approaches provide an alternative by means of credible intervals for each coefficient [17]. A low-dimensional projection estimator method was also designed to constructing confidence intervals [18]. A test statistic based on Lasso fitted values was also proposed [19]; 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 et al.* [12], whose procedure, detailed below, was later extended by *Meinshausen et al.* [13] using resampling and an aggregation of p-values for the controlling the FWER (FDR control still requires additional work).

B. The Original Screen and Clean Procedure

The screen and clean procedure is a two-stage method proposed by *Wasserman et al.* [12] to perform variable selection with statistical guarantees. The first stage screens variables to find a set of possibly relevant variables and the second stage cleans the set of candidate variables, thereby providing statistical guarantees on the risk of including irrelevant variables. The procedure considers a series of sparse models $\{\mathcal{F}_\lambda\}_{\lambda\in\Lambda}$, 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 $\mathcal{F}_\lambda$ is fitted to (part of) the data, thereby selecting a set of possibly relevant variables, also known as the support of the model $\mathcal{S}_\lambda \subseteq \{1,\ldots,p\}$. Then, in the second step, a model selection procedure chooses a single model $\mathcal{F}_{\hat\lambda}$ with its associated $\mathcal{S}_{\hat\lambda}$. Next, the cleaning stage eliminates possibly irrelevant variables from $\mathcal{S}_{\hat\lambda}$, resulting in the set $\hat{\mathcal{S}}$ that provably controls the type one error rate. The original procedure relies on three independent subsamples of the original data $\mathcal{D}$, so as to ensure the consistency of the overall process. The following chart summarizes this procedure, showing the actual use of data $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$ that is made at each step:

$$\overbrace{\{1,\ldots,p\} \xrightarrow[fit\ model]{step\,I(\mathcal{D}_1)} \{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} \xrightarrow[select\ model]{step\,II(\mathcal{D}_1,\mathcal{D}_2)} \mathcal{S}_{\hat\lambda}}^{screen}$$

$$\underbrace{\mathcal{S}_{\hat\lambda} \xrightarrow[test\ support]{step\,III(\mathcal{D}_3)} \hat{\mathcal{S}}}_{clean}$$

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 rigorous terms by *Meinshausen et al.* [13] as the “screening property” and “sparsity property”.

Empirically, *Wasserman et al.* [12] 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 always based on ordinary least squares (OLS) regression. The OLS parameters were adjusted on the “training” subsample $\mathcal{D}_1$, using the variables in $\{\mathcal{S}_\lambda\}_{\lambda\in\Lambda}$, and model selection consisted in minimizing the empirical error on the “validation” subsample $\mathcal{D}_2$ with respect to λ. Cleaning was then finally performed by testing the nullity of the OLS coefficients using the independent “test” subsample $\mathcal{D}_3$. *Wasserman et al.* [12] conclude that the variants using multivariate regression (Lasso and forward stepwise) have similar performances, way above univariate testing.

III. An Improved Screen and Clean Procedure

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

A. Adaptive-Ridge Cleaning Stage

The original cleaning stage of *Wasserman et al.* [12] 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 $|\mathcal{S}_{\hat\lambda}|$ that passed the screening stage is smaller than the number of examples $|\mathcal{D}_3|$ reserved for the cleaning stage;
- second, the only information retained from the screening stage is the support $\mathcal{S}_{\hat\lambda}$ 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 retaining the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ obtained at this stage. Our procedure allows for a cleaning stage in the high-dimensional setup (that is, $\left|\mathcal{S}_{\hat{\lambda}}\right| \gg |\mathcal{D}_3|$), and our experiments show that conveying the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ to the cleaning stage systematically improves the power of the procedure: the statistics produced result in dramatic increases in sensitivity (that is, in true positives) at any false discovery rate (see Figure 2).

Technically, the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ is forwarded to the cleaning stage via an adaptive-ridge penalty term. Adaptive refers here to the adaptation of the penalty terms to the data at hand. The penalty shape is adjusted to the "training" subsample $\mathcal{D}_1$, its strength is set thanks to the "validation" subsample $\mathcal{D}_2$, and it is finally applied to cleaning stage on $\mathcal{D}_3$. This process is detailed below.

1) Computing the Regression Coefficients: Our cleaning stage is specifically designed for a screening stage based on the Lasso or more generally on the Elastic-Net estimator [20], which is nowadays widely used to tackle simultaneously variable estimation and selection. The original Elastic-Net estimator is defined as:

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \Big(\lambda_1 \left\|\boldsymbol{\beta}\right\|_1 + \lambda_2 \left\|\boldsymbol{\beta}\right\|_2^2\Big) \quad , \tag{1}$$

where $\lambda = (\lambda_1, \lambda_2)$ is a two-dimensional 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

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

but, except for the numerical acceleration tricks mentioned at the end of Section III-A2, the overall feature selection process may be applied to any other form of $J(\boldsymbol{\beta})$, such as the ones used in classification.

Compared to the Lasso, the Elastic-Net requires the tuning of the additional hyper-parameter λ_2, thereby demanding more computations for model selection. This ℓ_2 penalization promotes stability, especially in the presence of correlations between features, while the solution remains sparse thanks to the ℓ_1 penalty [20]. In our framework, it also offers the possibility to select larger supports at the screening stage, thus favoring the "screening property" (more details will be given in Section III-B). We recall that selecting a large support is problematic when the cleaning stage relies on the OLS, which may then be unstable, or even ill-defined. We avoid this problem by using penalization at both stages of the feature selection method.

Our approach relies on an alternative view of the Elastic-Net, seen as an adaptive-ℓ_2 penalization scheme [21], [22]. This view is formalized by a variational form of the Elastic-Net:

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\tau}\in\mathbb{R}^p} \quad & J(\boldsymbol{\beta}) + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{\tau_j} + \lambda_2\right) \\ \text{s. t.} \quad & \sum_{j=1}^{p} \tau_j - \sum_{j=1}^{p} |\beta_j| \leq 0 \ , \ \tau_j \geq 0 \ , \ j = 1, \ldots, p \ . \end{aligned} \tag{2}$$

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:

$$\sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{|\hat{\beta}_j(\lambda)|} + \lambda_2\right) \quad , \tag{3}$$

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

Using this adaptive-ℓ_2 penalty returns the original Elastic-Net estimator. This equivalence will be used here for defining a data-dependent penalty, determined in the screening stage, that will also be applied in the cleaning stage. In this process, our primary aim is to retain the magnitude of the coefficients of $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ in addition to the support $\mathcal{S}_{\hat{\lambda}}$: the small coefficients of the screening stage will be encouraged to be small in the cleaning stage, whereas the largest ones will be less penalized. The expected side-effects of this penalization at the cleaning stage are to allow for the processing of more variables and to stabilize the estimation procedure.

Cleaning is eventually performed by testing the nullity of the adaptive-ridge coefficients using the independent "test" subsample $\mathcal{D}_3$. 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 Figure 2 of the numerical experiment section). However, using penalized estimators raises a difficulty for the calibration of the statistical tests derived from these statistics. We propose resolve this issue through the use of permutation tests.

2) Testing the Significance of the Regression 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.

A non-exact t-test can be severely off when the explanatory variables are strongly correlated [23]. For example, false positive rates as high as 32% have been reported for a significance level supposedly fixed at 5% [24]. 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

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

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 [25], the F-test is not exact for ridge regression, for the reasons already mentioned above (estimation bias and dependency between numerator and denominator. Here, we propose to approach the distribution of the F-statistic under the null hypothesis by randomization, thereby defining a permutation F-test.

Permutation tests are often used in a small sample setting where Gaussian approximations of the maximum likelihood estimates are not valid. To be exact, permutation tests assume some form of exchangeability. There is no finite-sample exact permutation test in multiple linear regression [26]. 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. Instead, we directly permute the values taken by the explicative variable to be tested, 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. Table IV shows that, compared to the standard t-test and F-test [25], 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 IV-A. Note that 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 [27]. The overall procedure is well calibrated as shown in Section IV.

Permutation tests rely on the fitting of several hundredth of randomized models. Appendix A details our efficient implementation that drastically reduces the computational cost.

B. *Modifications at Screening Stage*

Wasserman et al. [12] 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” [13] was established using the protocol described in Section II-B. To our knowledge, it remains to be proved for model selection based on cross-validation. However, *Wasserman et al.* [12] mention another procedure relying on two independent subsamples of the original data $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where model selection relies on leave-one-out cross-validation on $\mathcal{D}_1$ and $\mathcal{D}_2$ is reserved for cleaning. The following chart summarizes this modified procedure:

$$\overbrace{\{1,\ldots,p\} \xrightarrow[fit\ model]{step\, I(\mathcal{D}_1)} \{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} \xrightarrow[select\ model]{step\, II(\mathcal{D}_1)} \mathcal{S}_{\hat{\lambda}}}^{screen}$$

$$\underbrace{\mathcal{S}_{\hat{\lambda}} \xrightarrow[test\ support]{step\, III(\mathcal{D}_2)} \hat{\mathcal{S}}}_{clean}$$

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 Figure 1.

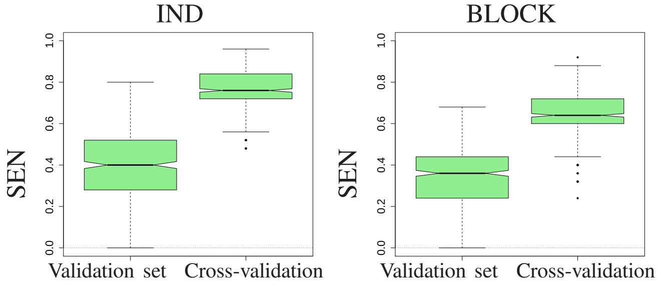

Fig. 1. Sensitivity of the screen and clean procedure (the higher, the better), for the two model selection strategies at the screening stage, and FDR controlled at 5% based on the permutation test. Lasso regression is used in the screening stage and adaptive-ridge regression in the cleaning stage. Each boxplot is computed based over 500 replications for the IND and BLOCK simulation designs, with $n = 250$, $p = 500$, $|\mathcal{S}^*| = 25$ and $\rho = 0.5$ (see Section IV-A for full description).

We slightly depart from *Wasserman et al.* [12], 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 et al.* [12], and later *Meinshausen et al.* [13] based model selection on the OLS estimate using the support $\mathcal{S}_\lambda$. This choice results in an implicit limitation of the size of the selected support $|\mathcal{S}_{\hat{\lambda}}| < \frac{n}{2}$, which is actually implemented more severely as $|\mathcal{S}_{\hat{\lambda}}| \leq \sqrt{n}$ and $|\mathcal{S}_{\hat{\lambda}}| \leq \frac{n}{6}$ by *Wasserman et al.* [12] and *Meinshausen et al.* [13] respectively. Our model selection criterion allows for more variables to be transferred at the cleaning stage, so that the screening property is more likely to hold.

IV. 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.

A. *Simulated Data*

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 $\mathcal{S}^\star = \left\{j \in \{1,...,p\} | \beta^\star_j \neq 0\right\}$ indexing its non-zero coefficients is small $|\mathcal{S}^\star| \ll p$.

1) Simulation Models: 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 $|\mathcal{S}^\star|$, the correlation structure of the explicative variables, the signal-to-noise ratio, and the relative magnitude of the relevant parameters $\{\beta^\star_j\}_{j\in\mathcal{S}^\star}$.

TABLE I. Expected false positive rate FPR (or type-I error) and sensitivity SEN (or power) computed over 500 simulations and over the variables selected in the screening stage by Lasso regression. The prescribed significance level is 5%. The IND, BLOCK, GROUP and TOEP$^-$ designs are fully described in Section IV-A.

Simulation 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

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

- IND independent explicative variables following a zero-mean, unit-variance Gaussian distribution: $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
- BLOCK dependent explicative variables following a zero-mean Gaussian distribution, with a block-diagonal covariance matrix: $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, 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 $\{1, ..., p\}$. This design is thus difficult for variable selection.
- GROUP same as BLOCK, except that the relevant variables are gathered a single block when $|\mathcal{S}^\star| = 25$ and in two blocks when $|\mathcal{S}^\star| = 50$, 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.

We fixed the signal-to-noise ratio $\boldsymbol{\beta}^{\star\top}\mathbf{\Sigma}\boldsymbol{\beta}^\star/\sigma^2 = 4$, which leads to feasible but challenging problems for model selection. Finally, the non-zero parameters $\beta_j^\star$ are drawn from a uniform distribution $\mathcal{U}(10^{-1}, 1)$ to enable different magnitudes.

In the following, we discuss the IND BLOCK, GROUP and TOEP$^-$ designs with $n = 250$, $p = 500$, $|\mathcal{S}^\star| = 25$ and $\rho = 0.5$, since the relative behavior of all methods is representative of the general pattern that we observed across all simulation settings.

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 can be thought as 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.

TABLE II. Average size of $\mathcal{S}_{\hat{\lambda}}$, computed over 500 simulations for each design, during the screening stage with Lasso and Elastic-Net regression.

Simulation design	IND	BLOCK	GROUP	TOEP$^-$
Lasso	96.95	89.51	39.77	75.80
Elastic net	97.11	91.72	58.40	86.26

TABLE III. False discovery rate FDR and sensitivity SEN, computed over 500 simulations for each design. The screening stage (*before cleaning*) is not calibrated; the cleaning stage (*after cleaning*) is calibrated to control the FDR below 5%.

Simulation design	IND		BLOCK		GROUP		TOEP$^-$	
	FDR	SEN	FDR	SEN	FDR	SEN	FDR	SEN
before cleaning	75.9	87.0	75.1	83.8	32.0	86.2	76.2	67.1
after cleaning	3.9	75.9	3.1	64.9	1.2	42.2	4.6	43.3

We conducted experiments with the Elastic-Net and the Lasso at the screening stage. Although they may result in different set sizes at the output of the screening stage (see Table II), they eventually end to identical results after the cleaning stage, in terms of FDR control and sensitivity. In what follows, we thus only report results with the simplest Lasso screening stage that relies on a single penalty parameter.

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 et al.* [12] and to the univariate approach. The variable selection method of *Lockhart et al.* [19] 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.

2) Importance of the Cleaning Stage: Table III 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 estimates are way too numerous: first, the penalty parameter is determined to optimize the cross-validated mean squared 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.

3) Comparisons of Controlled Selection Procedures: Figure 2 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-test for each variable independently, is by far the worst option in the IND and BLOCK designs, and by far the best in the GROUP design. In this last situation, the univariate

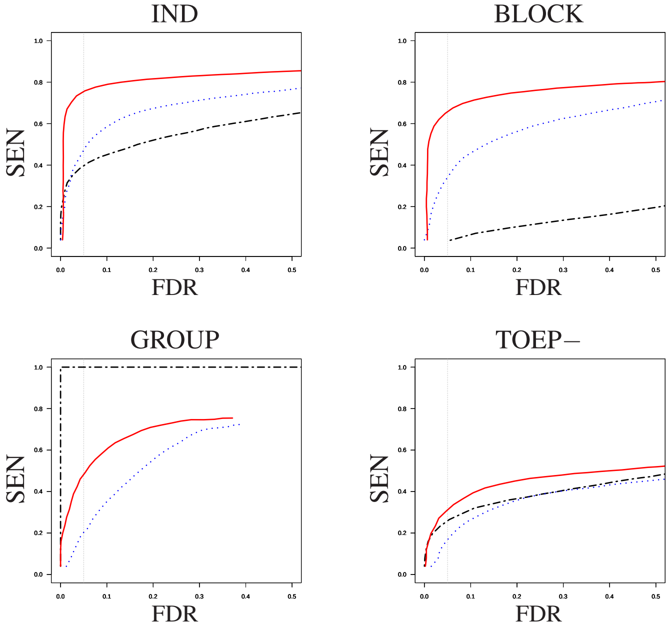

Fig. 2. Sensitivity (SEN) versus False Discovery Rate (FDR) (the higher, the better). Lasso screening with adaptive-ridge cleaning (red solid line), Lasso screening with OLS cleaning (blue dotted line) and univariate testing (black dot-dashed line). All curves are indexed by the rank of the test statistics, and averaged over the 500 simulations of each design. The vertical dotted line marks the 5% FDR level.

TABLE IV. FALSE DISCOVERY RATE FDR AND SENSITIVITY SEN, COMPUTED OVER 500 SIMULATIONS FOR EACH DESIGN. THE SCREENING STAGE (BEFORE CLEANING) IS NOT CALIBRATED; THE CLEANING STAGE IS CALIBRATED TO CONTROL THE FDR BELOW 5%, USING THE BENJAMINI-HOCHBERG PROCEDURE. OUR ADAPTIVE-RIDGE (AR) CLEANING IS COMPARED WITH THE ORIGINAL (OLS) CLEANING AND UNIVARIATE TESTING (UNIVAR).

Simulation design	IND		BLOCK		GROUP		TOEP$^-$	
	FDR	SEN	FDR	SEN	FDR	SEN	FDR	SEN
Before 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

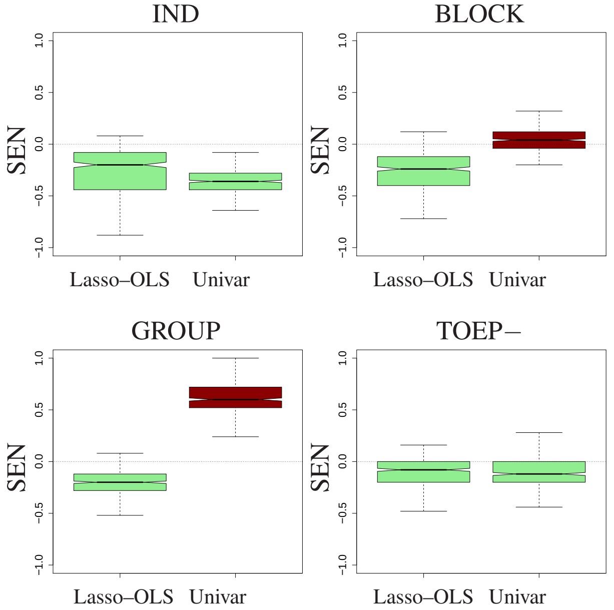

Fig. 3. Boxplots of sensitivity differences (the higher, the better) between our Lasso screening with adaptive-ridge regression cleaning, and (i) Lasso screening with OLS cleaning (Lasso–OLS), (ii) univariate testing (Univar). The statistics are computed over the 500 simulations of each design. The tests are calibrated to control the FDR below 5%. This constraint is always met for our method, and the boxes are colored light green when the FDR of the competitor is actually below 5%, and dark red otherwise (see Table IV).

approach confidently detects all the correlated variables of the relevant group and, 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: even when all relevant variables are correlated, like in the TOEP$^-$ design, univariate testing may not be the right choice. Finally, our adaptive-ridge cleaning systematically dominates the original OLS cleaning.

Table IV shows the actual operating conditions of the four 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 a 5% level, using a Benjamini-Hochberg correction. This control is always effective for all 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 comparable FDRs, while sensitivity gains range from 50% to 300%.

All differences in sensitivity are statistically significant. This is illustrated by the boxplots of Figure 3, which represent the *difference* of sensitivity between our Lasso-based screen and clean procedure and (i) its OLS couterpart, (ii) the univariate approach. Notches above zero indicate that the competitor has a significantly higher sensitivity, and dark red boxes indicate that its FDR is not properly controlled at 5%.

B. GWAS on HIV

We now compare the results of variable selection in a Genome Wide Association Study (GWAS) on HIV-1 infection [28]. 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

TABLE V. ADJUSTED p-VALUES (IN %) OBTAINED FROM THE BENJAMINI-HOCHBERG PROCEDURE FOR THE FOUR SNPS OF THE HIV DATA SELECTED AT A 25% FDR LEVEL. SCREENING IS PERFORMED BY LASSO, AND OUR ADAPTIVE-RIDGE CLEANING (LASSO–AR) IS COMPARED WITH THE ORIGINAL OLS CLEANING PROCEDURE (LASSO–OLS) AND WITH UNIVARIATE TESTING (UNIVAR).

SNP	Genomic Region	Lasso–AR	Lasso–OLS	Univar
rs10484554	MHC	2.9	21.9	0.003
rs2523619	MHC	5.8	97.0	0.2
rs2395029	MHC	9.7	62.0	1.3
rs6923486	other	13.1	17.9	99.5

quantitative response variable is the plasma HIV-RNA level, which is a marker of the HIV disease progression.

We used Lasso for screening, thereby selecting $|\mathcal{S}_{\hat{\lambda}}| = 29$ SNPs. Considering a 25% FDR level [28], the adaptive-ridge screening selects $|\hat{\mathcal{S}}| = 4$ SNPs as being associated with the plasma HIV-RNA, while OLS selects only $|\hat{\mathcal{S}}| = 2$ of them (see Table V). Among the 12 SNPs which were identified by *Dalmasso et al.* [28] from a univariate analysis in the MHC region, only 3 (RS10484554, RS2523619 and RS2395029) 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, those results are in line with the simulation study, where, 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.

V. DISCUSSION

We propose a new "screen and clean" procedure [12] for assessing the uncertainties pertaining to the selection of relevant variables in regression problems. Our main innovations lie in the cleaning stage. First, we use the connection between the Lasso and adaptive-ridge [21] to convey more information from the screening stage to the cleaning stage: the magnitude of the coefficients estimated at the screening stage is transferred to the cleaning stage through penalty parameters. We then propose a permutation test enabling the computation of p-values that are corrected for multiple testing [27].

Empirically, our procedure provides a better control of the False Discovery Rate, which extends to more 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. 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 [9], 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 cleaning stage can accommodate arbitrary penalties, and our efficient implementation applies to all penalties for which a quadratic variational formulation can be derived. This is particularly appealing for structured penalties such as the group-Lasso, allowing to use the knowledge of groups at the cleaning stage, through penalization coefficients.

On the theoretical side, many interesting issues are raised. In particular, we would like to back-up the empirical improvements that have been almost systematically observed by an apposite analysis. We believe that two tracks are promising: first by exploiting that the screening stage transfers a quantified response to the cleaning stage through the penalization coefficients, and second, that screening needs not to be stringent, due to the ability of our cleaning stage to handle more variables.

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.

REFERENCES

[1] Y. Wang, J. Yang, W. Yin, and W. Zhang, "A new alternating minimization algorithm for total variation image reconstruction," *SIAM J. Imaging Sciences*, vol. 1, no. 3, pp. 248–272, 2008.

[2] I. G. Chong and C. H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1–2, pp. 103–112, 2005.

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

[4] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.

[5] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, 2010.

[6] R. J. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

[8] N. Verzelen, "Minimax risks for sparse regressions: Ultra-high dimensional phenomenons," *Electronic Journal of Statistics*, vol. 6, pp. 38–90, 2012.

[9] P. Bühlmann, "Statistical significance in high-dimensional linear models," *Bernoulli*, vol. 19, pp. 1212–1242, 2013.

[10] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill, and V. Frouin, "Variable selection for generalized canonical correlation analysis," *Biostatistics*, vol. 15, no. 3, pp. 569–583, 2014.

[11] D. Balding, "A tutorial on statistical methods for population association studies," *Nature Reviews Genetics*, vol. 7, no. 10, pp. 781–791, october 2006.

[12] L. Wasserman and K. Roeder, "High-dimensional variable selection," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2178–2201, 2009.

[13] N. Meinshausen, L. Meier, and P. Bühlmann, "p-values for high-dimensional regression," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1671–1681, 2009.

[14] H. Liu and B. Yu, "Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression," *Electronic Journal of Statistics*, vol. 7, pp. 3124–3169, 2013.

[15] S. Dudoit and M. Van der Laan, *Multiple testing procedures with applications to genomics*. Springer, 2008.

[16] A. Chatterjee and S. N. Lahiri, "Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap," *The Annals of Statistics*, vol. 41, no. 3, pp. 1232–1259, 2013.

[17] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010.

[18] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, 2014.

[19] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani, "A significance test for the lasso," *The Annals of Statistics*, vol. 42, no. 2, pp. 413–468, 2014.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[21] Y. Grandvalet, "Least absolute shrinkage is equivalent to quadratic penalization," in *ICANN'98*, ser. Perspectives in Neural Computing, L. Niklasson, M. Bodén, and T. Ziemske, Eds., vol. 1. Springer, 1998, pp. 201–206.

[22] Y. Grandvalet and S. Canu, "Outcomes of the equivalence of adaptive ridge with least absolute shrinkage," in *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds. MIT Press, 1999, pp. 445–451.

[23] A. M. Halawa and M. Y. El Bassiouni, "Tests of regressions coefficients under ridge regression models," *Journal of Statistical Computation and Simulation*, vol. 65, no. 1, pp. 341–356, 1999.

[24] E. Cule, P. Vineis, and M. De Lorio, "Significance testing in ridge regression for genetic data," *BMC Bioinformatics*, vol. 12, no. 372, pp. 1–15, 2011.

[25] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall, 1990, vol. 43.

[26] M. J. Anderson and J. Robinson, "Permutation tests for linear models," *Australian & New Zealand Journal of Statistics*, vol. 43, no. 1, pp. 75–88, 2001.

[27] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[28] C. Dalmasso, W. Carpentier, L. Meyer, C. Rouzioux, C. Goujard, M.-L. Chaix, O. Lambotte, V. Avettand-Fenoel, S. Le Clerc, L. Denis de Senneville, C. Deveau, F. Boufassa, P. Debre, J.-F. Delfraissy, P. Broet, and I. Theodorou, "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*, vol. 3, no. 12, p. e3907, 2008.

Appendix A
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 $q = |\mathcal{S}_{\hat{\lambda}}|$ 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 $O(Bq(q^3 + q^2 n))$. In the situation where $q \ll B$, great computing savings can be obtained using block-wise decompositions and inversions.

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 $(\lambda_1/\tau_j^\star + \lambda_2)$, 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: $\mathbf{X}^{(b)} = (\mathbf{x}_j^{(b)}, \mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, ...\mathbf{x}_p)$. Then, $\hat{\boldsymbol{\beta}}^{(b)}$ can be efficiently computed by using $a^{(b)} \in \mathbb{R}$, $\mathbf{X}_{\backslash j} \in \mathbb{R}^{q-1}$ and $\hat{\boldsymbol{\beta}}_{\backslash j} \in \mathbb{R}^{q-1}$ defined as follows:

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

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

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

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

$$(\mathbf{X}_{\backslash j}^\top \mathbf{X}_{\backslash j} + \boldsymbol{\Lambda}_{\backslash j})^{-1} = \left[\left(\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\backslash j \backslash j} - \\ \left[\left(\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\backslash j j} \left[\left(\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{jj}^{-1} \left[\left(\mathbf{X}^\top \mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{j \backslash j}$$

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.

Appendix B
Software

Software and simulations are in the form of R package named "ridgeAdap" available on the personal author page (https://www.hds.utc.fr/∼becujean/dokuwiki/doku.php).