SLOPE – Adaptive Variable Selection via Convex Optimization

Małgorzata Bogdan[a] Ewout van den Berg[b] Chiara Sabatti[c,d] Weijie Su[d]
Emmanuel J. Candès[d,e*]

June 2015[†]

[a] Department of Mathematics, Wrocław University of Technology, Poland
[b] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.
[c] Department of Health Research and Policy, Stanford University, Stanford CA 94305, U.S.A.
[d] Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
[e] Department of Mathematics, Stanford University, Stanford, CA 94305, U.S.A.

Abstract

We introduce a new estimator for the vector of coefficients β in the linear model $y = X\beta + z$, where X has dimensions $n \times p$ with p possibly larger than n. SLOPE, short for Sorted L-One Penalized Estimation, is the solution to

$$\min_{b\in\mathbb{R}^p} \quad \tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1|b|_{(1)} + \lambda_2|b|_{(2)} + \cdots + \lambda_p|b|_{(p)},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$ are the decreasing absolute values of the entries of b. This is a convex program and we demonstrate a solution algorithm whose computational complexity is roughly comparable to that of classical ℓ_1 procedures such as the Lasso. Here, the regularizer is a sorted ℓ_1 norm, which penalizes the regression coefficients according to their rank: the higher the rank—i. e. the stronger the signal—the larger the penalty. This is similar to the Benjamini and Hochberg (1995) procedure (BH), which compares more significant p-values with more stringent thresholds. One notable choice of the sequence $\{\lambda_i\}$ is given by the BH critical values $\lambda_{\mathrm{BH}}(i) = z(1-i\cdot q/2p)$, where $q \in (0,1)$ and $z(\alpha)$ is the quantile of a standard normal distribution. SLOPE aims to provide finite sample guarantees on the selected model; of special interest is the false discovery rate (FDR), defined as the expected proportion of irrelevant regressors among all selected predictors. Under orthogonal designs, SLOPE with λ_{BH} provably controls FDR at level q. Moreover, it also appears to have appreciable inferential properties under more general designs X while having substantial power, as demonstrated in a series of experiments running on both simulated and real data.

Keywords. Sparse regression, variable selection, false discovery rate, Lasso, sorted ℓ_1 penalized estimation (SLOPE).

Introduction

Analyzing and extracting information from datasets where the number of observations n is smaller than the number of variables p is one of the challenges of the present "big-data" world. In response, the statistics literature of the past two decades documents the development of a variety

*Corresponding author

[†]An earlier version of the paper appeared on arXiv.org in October 2013: arXiv:1310.1969v2

of methodological approaches to address this challenge. A frequently discussed problem is that of linking, through a linear model, a response variable y to a set of predictors $\{X_j\}$ taken from a very large family of possible explanatory variables. In this context, the Lasso (Tibshirani, 1996) and the Dantzig selector (Candès and Tao, 2007), for example, are computationally attractive procedures offering some theoretical guarantees, and with consequent wide-spread application. In spite of this, there are some scientific problems where the outcome of these procedures is not entirely satisfying, as they do not come with a machinery allowing us to make inferential statements on the validity of selected models in finite samples. To illustrate this, we resort to an example.

Consider a study where a geneticist has collected information about n individuals by having identified and measured all p possible genetics variants in a genomic region. The geneticist wishes to discover which variants cause a certain biological phenomenon, such as an increase in blood cholesterol level. Measuring cholesterol levels in a new individual is cheaper and faster than scoring his or her genetic variants, so that predicting y in future samples given the value of the relevant covariates is not an important goal. Instead, correctly identifying functional variants is relevant. A genetic polymorphism *correctly* implicated in the determination of cholesterol levels points to a specific gene and to a biological pathway that might not be previously known to be related to blood lipid levels and, therefore, promotes an increase in our understanding of biological mechanisms, as well as providing targets for drug development. On the other hand, the *erroneous* discovery of an association between a genetic variant and cholesterol levels will translate to considerable waste of time and money, which will be spent in trying to verify this association with direct manipulation experiments. It is worth emphasizing that some of the genetic variants in the study have a biological effect while others do not—there is a ground truth that statisticians can aim to discover. To be able to share the results with the scientific community in a convincing manner, the researcher needs to be able to attach some finite sample confidence statements to his/her findings. In a more abstract language, our geneticist would need a tool that privileges correct model selection over minimization of prediction error, and would allow for inferential statements to be made on the validity of his/her selections. This paper presents a new methodology that attempts to address some of these needs.

We imagine that the n-dimensional response vector y is truly generated by a linear model of the form

$$y = X\beta + z,$$

with X an $n \times p$ design matrix, β a p-dimensional vector of regression coefficients and z an $n \times 1$ vector of random errors. We assume that all relevant variables (those with $\beta_i \neq 0$) are measured in addition to a large number of irrelevant ones. As any statistician knows, these assumptions are quite restrictive, but they are a widely accepted starting point. To formalize our goal, namely, the selection of important variables accompanied by a finite sample confidence statement, we seek a procedure that controls the expected proportion of irrelevant variables among the selected. In a scientific context where selecting a variable corresponds to making a discovery, we aim at controlling the False Discovery Rate (FDR). The FDR is of course a well-recognized measure of global error in multiple testing and effective procedures to control it are available: indeed, the Benjamini and Hochberg (1995) procedure (BH) inspired the present proposal. The connection between multiple testing and model selection has been made before (see e.g., Bauer et al. (1988), Foster and George (1994), Abramovich and Benjamini (1995), Abramovich et al. (2006), Bogdan et al. (2008)) and others in recent literature have tackled the challenges encountered by our geneticists: we will discuss the differences between our approach and others in later sections as appropriate. The procedure we introduce in this paper is, however, entirely new. Variable selection is achieved by solving a convex

problem not previously considered in the statistical literature, and which marries the advantages of ℓ_1 penalization with the adaptivity inherent in strategies like BH.

Section 1 of this paper introduces SLOPE, our novel penalization strategy, motivates its construction in the context of orthogonal designs, and places it in the context of current knowledge of effective model selection strategies. Section 2 describes the algorithm we developed and implemented to find SLOPE estimates. Section 3 showcases the application of our novel procedure in a variety of settings: we illustrate how it effectively solves a multiple testing problem with positively correlated test statistics; we discuss how regularizing parameters should be chosen in non-orthogonal designs; we investigate the robustness of SLOPE to some violations of model assumptions and we apply it to a genetic dataset, not unlike our idealized example. Section 4 concludes the paper with a discussion comparing our methodology to other recently introduced proposals as well as outlining open problems.

1 Sorted L-One Penalized Estimation (SLOPE)

1.1 Adaptive penalization and multiple testing in orthogonal designs

To build intuition behind SLOPE, which encompasses our proposal for model selection in situations where $p > n$, we begin by considering the case of orthogonal designs and i.i.d. Gaussian errors with known standard deviation, as this makes the connection between model selection and multiple testing natural. Since the design is orthogonal, $X'X = I_p$, and the regression $y = X\beta + z$ with $z \sim \mathcal{N}(0, \sigma^2 I_n)$ can be recast as

$$\tilde{y} = X'y = X'X\beta + X'z = \beta + X'z \sim \mathcal{N}(\beta, \sigma^2 I_p). \tag{1.1}$$

In some sense, the problem of selecting the correct model reduces to the problem of testing the p hypotheses $H_{0,j} : \beta_j = 0$ versus two sided alternatives $H_{1,j} : \beta_i \neq 0$. When p is large, a multiple comparison correction strategy is called for and we consider two popular procedures.

- *Bonferroni's method.* To control the familywise error rate[1] (FWER) at level $\alpha \in [0, 1]$, one can apply Bonferroni's method, and reject $H_{0,j}$ if $|\tilde{y}_j|/\sigma > \Phi^{-1}(1 - \alpha/2p)$, where $\Phi^{-1}(\alpha)$ is the αth quantile of the standard normal distribution. Hence, Bonferroni's method defines a comparison threshold that depends only on the number of covariates, p, and the noise level.

- *Benjamini-Hochberg step-up procedure.* To control the FDR at level $q \in [0, 1]$, BH begins by sorting the entries of $\tilde{y}$ in decreasing order of magnitude, $|\tilde{y}|_{(1)} \geq |\tilde{y}|_{(2)} \geq \cdots \geq |\tilde{y}|_{(p)}$, which yields corresponding ordered hypotheses $H_{(1)}, \ldots, H_{(p)}$. (Note that here, as in the rest of the paper, (1) indicates the largest element of a set, instead of the smallest. This breaking with common convention allows us to keep (1) as the index for the most 'interesting' hypothesis). Then BH rejects all hypotheses $H_{(i)}$ for which $i \leq i_{\text{BH}}$, where i_{BH} is defined by

  $$i_{\text{BH}} = \max\{i : |\tilde{y}|_{(i)}/\sigma \geq \Phi^{-1}(1 - q_i)\}, \quad q_i = i \cdot q/2p \tag{1.2}$$

  (with the convention that $i_{\text{BH}} = 0$ if the set above is empty). Letting V (resp. R) be the total number of false rejections (resp. total number of rejections), Benjamini and Hochberg (1995) showed that for BH

  $$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = q\frac{p_0}{p}, \tag{1.3}$$

[1] Recall that the FWER is the probability of at least one false rejection.

where p_0 is the number of true null hypotheses, $p_0 := |\{i : \beta_i = 0\}| = p - \|\beta\|_{\ell_0}$.

In contrast to Bonferroni's method, BH is an adaptive procedure in the sense that the threshold for rejection $|y|_{(i_{\mathrm{BH}})}$ is defined in a data-dependent fashion, and is sensitive to the sparsity and magnitude of the true signals. In a setting where there are many large β_j's, the last selected variable needs to pass a far less stringent threshold than it would in a situation where no β_j is truly different from 0. It has been shown in a variety of papers (see e.g., Abramovich et al. (2006); Bogdan et al. (2011); Wu and Zhou (2013); Frommlet and Bogdan (2013)) that this behavior allows BH to adapt to the unknown signal sparsity, resulting in some important asymptotic optimality properties.

We now consider how the Lasso would behave in this setting. The solution to

$$\min_{b\in\mathbb{R}^p} \tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1} \tag{1.4}$$

in the case of orthogonal designs is given by soft-thresholding. In particular, the Lasso estimate $\hat{\beta}_j$ is not zero if and only if $|\tilde{y}_j| > \lambda$. That is, variables are selected using a non-adaptive threshold λ. Mindful of the costs associated with the selection of irrelevant variables, we can control the FWER by setting $\lambda_{\mathrm{Bonf}} = \sigma \cdot \Phi^{-1}(1 - \alpha/2p) \approx \sigma \cdot \sqrt{2 \log p}$.[2] This choice, however, is likely to result in a loss of power, and may not strike the right balance between errors of type I and missed discoveries. Choosing a value of λ substantially smaller than λ_{Bonf} in a non-data dependent fashion, would lead not only to a loss of FWER control, but also of FDR control since FDR and FWER are identical measures under the global null in which all our variables are irrelevant. Another strategy is to use cross-validation. However, this data-dependent approach for selecting the regularization parameter λ targets the minimization of prediction error, and does not offer guarantees with respect to model selection (see Section 1.3.3). Our idea to achieve adaptivity, thereby increasing power while controlling some form of type-one error is to break the monolithic penalty $\lambda\|\beta\|_{\ell_1}$, which treats every variable in the same manner. Set

$$\lambda_{\mathrm{BH}}(i) \stackrel{\mathrm{def}}{=} \Phi^{-1}(1 - q_i), \quad q_i = i \cdot q/2p,$$

and consider the following program

$$\min_{b\in\mathbb{R}^p} \tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sigma \cdot \sum_{i=1}^{p} \lambda_{\mathrm{BH}}(i)|b|_{(i)}, \tag{1.5}$$

where $|b|_{(1)} \geq |b|_{(2)} \geq \cdots \geq |b|_{(p)}$ are the order statistics of the absolute values of the coordinates of b: in equation (1.5) different variables receive different levels of penalization depending on their relative importance. While the similarities of equation (1.5) with BH are evident, the solution to equation (1.5) is not a series of scalar thresholding operations: the procedures are not—even in this case of orthogonal variables–exactly equivalent. Nevertheless, an upper bound on FDR proved in the supplementary appendix (Bogdan et al., 2015) can still be assured:

Theorem 1.1. *In the linear model with orthogonal design X and $z \sim \mathcal{N}(0, \sigma^2 I_n)$, the procedure equation* (1.5) *rejecting hypotheses for which $\hat{\beta}_j \neq 0$, has an FDR obeying*

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] \leq q\frac{p_0}{p}. \tag{1.6}$$

[2]For large t, we have $1 - \Phi(t) = t^{-1}\phi(t)(1 + o(t^{-1}))$,where $\phi(\cdot)$ denotes the density of $N(0, 1)$. Our approximation comes from setting the right-hand side to $\alpha/2p$ for a fixed value of α, say $\alpha = 0.05$, and a large value of p.

Figure 1 illustrates the FDR achieved by equation (1.5) in simulations using a $5,000 \times 5,000$ orthogonal design X and nonzero regression coefficients equal to $5\sqrt{2 \log p}$.

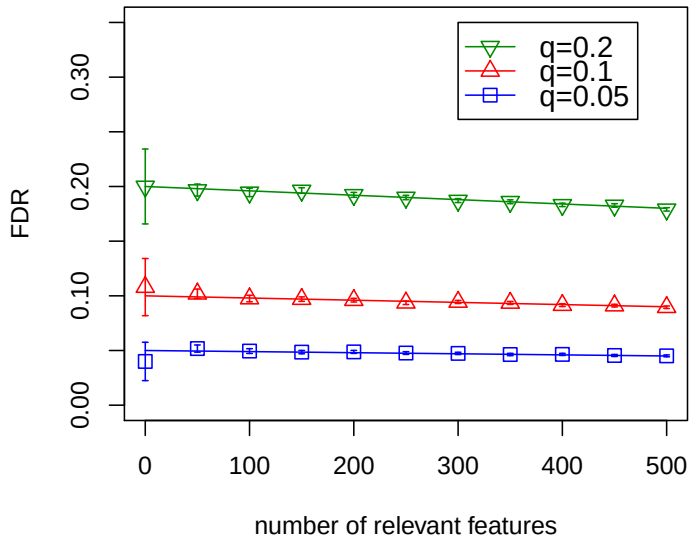


Figure 1: FDR of equation (1.5) in an orthogonal setting in which $n = p = 5000$. Straight lines correspond to $q \cdot p_0/p$, marked points indicate the average False Discovery Proportion (FDP) across 500 replicates, and bars correspond to $\pm$ 2 SE.

We conclude this section with several remarks describing the properties of our procedure under orthogonal designs.

1. While the $\lambda_{\mathrm{BH}}(i)$'s are chosen with reference to BH, equation (1.5) is neither equivalent to the step-up procedure described above nor to the step-down version.[3]

2. The proposal equation (1.5) is sandwiched between the step-down and step-up procedures in the sense that it rejects at most as many hypotheses as the step-up procedure and at least as many as the step-down cousin, also known to control the FDR (Sarkar, 2002).

3. The fact that equation (1.5) controls FDR is not a trivial consequence of this sandwiching.

The observations above reinforce the fact that equation (1.5) is different from the procedure known as *FDR thresholding* developed by Abramovich and Benjamini (1995) in the context of wavelet estimation and later analyzed in Abramovich et al. (2006). With $t_{\mathrm{FDR}} = |\tilde{y}|_{(i_{\mathrm{BH}})}$, FDR thresholding sets

$$\hat{\beta}_i = \begin{cases} \tilde{y}_i & |\tilde{y}_i| \geq t_{\mathrm{FDR}} \\ 0 & |\tilde{y}_i| < t_{\mathrm{FDR}}. \end{cases} \tag{1.7}$$

This is a hard-thresholding estimate but with a data-dependent threshold: the threshold decreases as more components are judged to be statistically significant. It has been shown that this simple estimate is asymptotically minimax throughout a range of sparsity classes (Abramovich et al.,

[3]The step-down version rejects $H_{(1)}, \ldots, H_{(i-1)}$, where i is the first time at which $|\tilde{y}_i|/\sigma \leq \Phi^{-1}(1 - q_i)$.

2006). Our method is similar in the sense that it also chooses an adaptive threshold reflecting the BH procedure. However, it does not produce a hard-thresholding estimate. Rather, owing to nature of the sorted ℓ_1 norm, it outputs a sort of soft-thresholding estimate. A substantial difference is that FDR thresholding equation (1.7) is designed specifically for orthogonal designs, whereas the formulation (1.5) can be employed for arbitrary design matrices leading to efficient algorithms. Aside from algorithmic issues, the choice of the λ sequence is, however, generally challenging.

1.2 SLOPE

While orthogonal designs have helped us define the program equation (1.5), this penalized estimation strategy is clearly applicable in more general settings. To make this explicit, it is useful to introduce the *sorted ℓ_1 norm*: letting $\lambda \neq 0$ be a nonincreasing sequence of nonnegative scalars,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0, \tag{1.8}$$

we define the sorted-ℓ_1 norm of a vector $b \in \mathbb{R}^p$ as[4]

$$J_\lambda(b) = \lambda_1 |b|_{(1)} + \lambda_2 |b|_{(2)} + \cdots + \lambda_p |b|_{(p)}. \tag{1.9}$$

Proposition 1.2. *The functional equation* (1.9) *is a norm provided equation* (1.8) *holds.*

The proof of Proposition 1.2 is provided in the supplementary appendix (Bogdan et al., 2015). Now define SLOPE as the solution to

$$\text{minimize} \quad \tfrac{1}{2}\|y - Xb\|^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}. \tag{1.10}$$

As a convex program, SLOPE is tractable: as a matter of fact, we shall see in Section 2 that its computational cost is roughly the same as that of the Lasso. Just as the sorted ℓ_1 norm is an extension of the ℓ_1 norm, SLOPE can be also viewed as an extension of the Lasso. SLOPE's general formulation, however, allows to achieve the adaptivity we discussed earlier. The case of orthogonal regressors suggests one particular choice of a λ sequence and we will discuss others in later sections.

1.3 Relationship to other model selection strategies

Our purpose is to bring the program equation (1.10) to the attention of the statistical community: this is a computational tractable proposal for which we provide robust algorithms; it is very similar to BH when the design is orthogonal; and has promising properties in terms of FDR control for general designs. We now compare it with two other commonly used approaches to model selection: methods based on the minimization of ℓ_0 penalties and the adaptive lasso. We discuss these here because they allow us to emphasize the motivation and characteristics of the SLOPE algorithm. We also note that the last few years have witnessed a substantive push towards the development of an inferential framework after selection (see e.g., Benjamini and Yekutieli (2005); Wasserman and Roeder (2009); Meinshausen et al. (2009); Berk et al. (2013); Zhang and Zhang (2013); van de Geer et al. (2014); Meinshausen and Bühlmann (2010); Efron (2011); Javanmard

[4]Observe that when all the λ_i's take on an identical positive value, the sorted ℓ_1 norm reduces to the usual ℓ_1 norm (up to a multiplicative factor). Also, when $\lambda_1 > 0$ and $\lambda_2 = \cdots = \lambda_p = 0$, the sorted ℓ_1 norm reduces to the ℓ_∞ norm (again, up to a multiplicative factor).

and Montanari (2013a,b); Bühlmann (2013); Lockhart et al. (2014)), with the exploration of quite different viewpoints. We will comment on the relationships between SLOPE and some of these methods, developed while editing this work, in the discussion section.

1.3.1 Methods based on ℓ_0 penalties

Canonical model selection procedures find estimates $\hat{\beta}$ by solving

$$\min_{b\in\mathbb{R}^p} \|y - Xb\|_{\ell_2}^2 + \lambda\|b\|_{\ell_0}, \tag{1.11}$$

where $\|b\|_{\ell_0}$ is the number of nonzero components in b. The idea behind such procedures is to achieve the best possible trade-off between the goodness of fit and the number of variables included in the model. Popular selection procedures such as AIC (Akaike, 1974) and C_p (Mallows, 1973) are of this form: when the errors are i.i.d. $\mathcal{N}(0,\sigma^2)$, AIC and C_p take $\lambda = 2\sigma^2$. In the high-dimensional regime, such a choice typically leads to including very many irrelevant variables, yielding rather poor predictive properties when the true vector of regression coefficients is sparse. In part to remedy this problem, Foster and George (1994) developed the risk inflation criterion (RIC): they proposed using a larger value of λ, effectively proportional to $2\sigma^2 \log p$, where p is the total number of variables in the study. Under orthogonal designs, if we associate nonzero fitted coefficients with rejections, this yields FWER control. Unfortunately, RIC is also rather conservative and, therefore, it may not have much power in detecting variables with nonvanishing regression coefficients unless they are very large.

The above dichotomy has been recognized for some time now and several researchers have proposed more adaptive strategies. One frequently discussed idea in the literature is to let the parameter λ in equation (1.11) decrease as the number of included variables increases. For instance, when minimizing

$$\|y - Xb\|_{\ell_2}^2 + p(\|b\|_{\ell_0}),$$

penalties with appealing information- and decision-theoretic properties are roughly of the form

$$p(k) = 2\sigma^2 k \log(p/k) \quad \text{or} \quad p(k) = 2\sigma^2 \sum_{1\le j\le k} \log(p/j). \tag{1.12}$$

Among others, we refer the interested reader to Foster and Stine (1999); Birgé and Massart (2001) and to Tibshirani and Knight (1999) for related approaches.

Interestingly, for large p and small k these penalties are close to the FDR related penalty

$$p(k) = \sigma^2 \sum_{1\le j\le k} \lambda_{\mathrm{BH}}^2(i), \tag{1.13}$$

proposed in Abramovich et al. (2006) in the context of the estimation of the vector of normal means, or regression under the orthogonal design (see the preceding section) and further explored in Benjamini and Gavrilov (2009). Due to an implicit control of the number of false discoveries, similar model selection criteria are appealing in gene mapping studies (see e.g., Frommlet et al. (2012)).

The problem with these selection strategies is that, in general, they are computationally intractable. Solving (1.12) would involve a brute-force search essentially requiring to fit least-squares

estimates for *all* possible subsets of variables. This is not practical for even moderate values of p, e.g., for $p > 60$.

The decaying sequence of the smoothing parameters in SLOPE goes along the line of the adaptive ℓ_0 penalties specified in equation (1.12), in which the 'cost per variable included' decreases as more get selected. However, SLOPE is computationally tractable and can be easily evaluated even for large-dimensional problems.

1.3.2 Adaptive Lasso

Perhaps the most popular alternative to the computationally intractable ℓ_0 penalization methods is the Lasso. We have already discussed some of the limitations of this approach with respect to FDR control and now wish to explore further the connections between SLOPE and variants of this procedure. It is well known that the Lasso estimates of the regression coefficients are biased due to the shrinkage imposed by the ℓ_1 penalty. To increase the accuracy of the estimation of large signals and eliminate some false discoveries the adaptive or reweighted versions of Lasso were introduced (see e.g., Zou (2006) or Candès et al. (2008)). In these procedures the smoothing parameters $\lambda_1, \ldots, \lambda_p$ are adjusted to the unknown signal magnitudes based on some estimates of regression coefficients, perhaps obtained through previous iterations of Lasso. The idea is then to consider a weighted penalty $\sum_i w_i|b_i|$, where w_i is inversely proportional to the estimated magnitudes so that large regression coefficients are shrunk less than smaller ones. In some circumstances, such adaptive versions of Lasso outperform its regular version (Zou, 2006).

The idea behind SLOPE is entirely different. In the adaptive Lasso, the penalty tends to decrease as the magnitude of coefficients increases. In our approach, the exact opposite happens. This comes from the fact that we seek to adapt to the unknown signal sparsity and control FDR. As shown in Abramovich et al. (2006), FDR controlling properties can have interesting consequences for estimation. In practice, since the SLOPE sequence $\lambda_1 \geq \cdots \geq \lambda_p$ leading to FDR control is typically rather large, we do not recommend using SLOPE directly for the estimation of regression coefficients. Instead we propose the following two-stage procedure: in the first step, SLOPE is used to identify significant predictors; in the second step, the corresponding regression coefficients are estimated using the least squares method within the identified sparse regression model. Such a two-step procedure, previously proposed in the context of LASSO (see e.g. Meinshausen (2007)), can be thought of as an extreme case of reweighting, where the selected variables are not penalized while those that are not selected receive an infinite penalty. As shown below, these estimates have very good properties when the coefficient sequence β is sparse.

1.3.3 A first illustrative simulation

To concretely illustrate the specific behavior of SLOPE compared to more traditional penalized approaches, we rely on the simulation of a relatively simple data structure. We set $n = p = 5000$ and generate the entries of the design matrix with i.i.d. $\mathcal{N}(0, 1/n)$ entries. The number of true signals k varies between 0 and 50 and their magnitudes are set to $\beta_i = \sqrt{2 \log p} \approx 4.1$, while the variance of the error term is assumed known and equal to 1. Since the expected value of the maximum of p independent standard normal variables is approximately equal to $\sqrt{2 \log p}$ and that the whole distribution of the maximum concentrates around this value, this choice of model parameters makes the sparse signal barely distinguishable from the noise because the nonzero means are at the level of the largest null statistics. We refer to e.g. Ingster (1999) for a precise discussion

of the limits of detectability in sparse mixtures.

We fit these observations with three procedures: 1) Lasso with parameter $\lambda_{\text{Bonf}} = \sigma \cdot \Phi^{-1}(1 - \alpha/2p)$, which controls FWER weakly; 2) Lasso with the smoothing parameter λ_{CV} chosen with 10-fold cross-validation; 3) SLOPE with a sequence $\lambda_1, \ldots, \lambda_p$ defined in Section 3.2.2, expression equation (3.8). The level α for λ_{Bonf} and q for FDR control in SLOPE are both set to 0.1. To compensate for the fact that lasso with λ_{Bonf} and SLOPE tend to apply a much more stringent penalization than lasso with λ_{CV}—which aims to minimize prediction error—we have "de-biased" their resulting $\hat{\beta}$, using ordinary least squares to estimate the coefficients of the variables selected by lasso–λ_{Bonf} and SLOPE (see Meinshausen (2007)).

We compare the procedures on the basis of three criteria: a) FDR, b) power, c) relative squared error $\|X\hat{\beta} - X\beta\|_{\ell^2}^2 / \|X\beta\|_{\ell_2}^2$. Note that only the first of these measures is meaningful for the case where $k = 0$, and in such case FDR=FWER.

Figure 2 reports the results of 500 independent replicates. The three approaches exhibit quite dramatically different properties with respect to model selection. SLOPE controls FDR at the desired level 0.1 for the explored range of k; as k increases, its power goes from 45% to 70%. Lasso–λ_{Bonf} has FDR =0.1 at $k = 0$, and a much lower one for the remaining values of k; this results in a loss of power with respect to SLOPE: irrespective of k, the power is less than 45%. Cross-validation chooses a λ that minimizes an estimate of prediction error, and in our experiments, λ_{CV} is quite smaller than a penalization parameter chosen with FDR control in mind. This results in greater power than SLOPE, but with a much larger FDR (80% on average).

Figure 2(c) illustrates the relative mean-square error, which serves as a measure of prediction accuracy. It is remarkable how, despite the fact that Lasso–λ_{CV} has higher power, SLOPE builds a better predictive model since it has a lower prediction error percentage for all the sparsity levels considered.

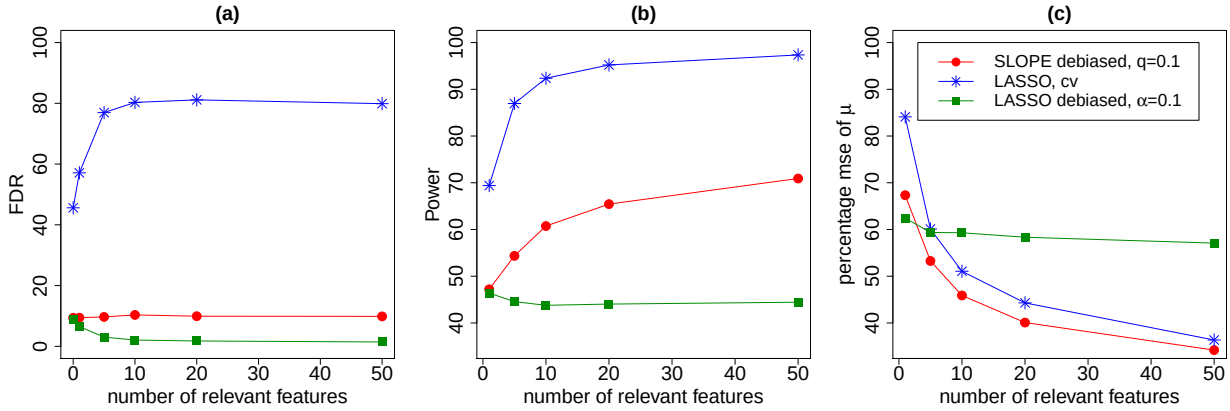


Figure 2: Properties of different procedures as a function of the true number of nonzero regression coefficients: (a) FDR, (b) Power, (c) Relative MSE defined as the average of $100 \cdot \|\hat{\mu} - \mu\|_{\ell_2}^2 / \|\mu\|_{\ell_2}^2$, with $\mu = X\beta$, $\hat{\mu} = X\hat{\beta}$. The design matrix entries are i.i.d. $\mathcal{N}(0, 1/n)$, $n = p = 5,000$, all nonzero regression coefficients are equal to $\sqrt{2 \log p} \approx 4.13$, and $\sigma^2 = 1$. Each point in the figures corresponds to the average of 500 replicates.

2 Algorithms

In this section, we present effective algorithms for computing the solution to SLOPE equation (1.10), which rely on the numerical evaluation of the proximity operator (prox) to the sorted ℓ_1 norm.

2.1 Proximal gradient algorithms

SLOPE is a convex optimization problem of the form

$$\text{minimize} \quad f(b) = g(b) + h(b), \tag{2.1}$$

where g is smooth and convex, and h is convex but not smooth. In SLOPE, g is the residual sum of squares and, therefore, quadratic while h is the sorted ℓ_1 norm. A general class of algorithms for solving problems of this kind are known as *proximal gradient methods*, see Nesterov (2007); Parikh and Boyd (2013) and references therein. These are iterative algorithms operating as follows: at each iteration, we hold a guess b of the solution and compute a local approximation to the smooth term g of the form

$$g(b) + \langle \nabla g(b), x - b \rangle + \frac{1}{2t}\|x - b\|_{\ell_2}^2.$$

This is interpreted as the sum of a Taylor approximation of g and of a proximity term; as we shall see, this term is responsible for searching an update reasonably close to the current guess b, and t can be thought of as a step size. Then the next guess b_+ is the unique solution to

$$\begin{aligned} b_+ &= \arg\min_x \quad \left\{ g(b) + \langle \nabla g(b), x - b \rangle + \frac{1}{2t}\|x - b\|_{\ell_2}^2 + h(x) \right\} \\ &= \arg\min_x \quad \left\{ \frac{1}{2t}\|(b - t\nabla g(b)) - x\|_{\ell_2}^2 + h(x) \right\} \end{aligned}$$

(unicity follows from strong convexity). In the literature, the mapping

$$x(y) = \arg\min_x \quad \left\{ \frac{1}{2t}\|y - x\|_{\ell_2}^2 + h(x) \right\}$$

is called the proximal mapping or prox for short, and denoted by $x = \text{prox}_{th}(y)$.

The prox of the ℓ_1 norm is given by entry-wise soft-thresholding (Parikh and Boyd, 2013, page 150) so that a proximal gradient method to solve the Lasso would take the following form: starting with $b^0 \in \mathbb{R}^p$, inductively define

$$b^{k+1} = \eta_{\lambda t_k}(b^k - t_k X'(Xb^k - y); t_k\lambda),$$

where $\eta_\lambda(y) = \text{sign}(y) \cdot (|y| - \lambda)_+$ and $\{t_k\}$ is a sequence of step sizes. Hence, we can solve the Lasso by iterative soft thresholding.

It turns out that one can compute the prox to the sorted ℓ_1 norm in nearly the same amount of time as it takes to apply soft thesholding. In particular, assuming that the entries are sorted (an order $p \log p$ operation), we shall demonstrate a linear-time algorithm. Hence, we may consider a proximal gradient method for SLOPE as in Algorithm 1.

Algorithm 1 Proximal gradient algorithm for SLOPE equation (1.10)

Require: $b^0 \in \mathbb{R}^p$

1: **for** $k = 0, 1, \ldots$ **do**
2: $\quad b^{k+1} = \mathrm{prox}_{t_k J_\lambda}(b^k - t_k X'(Xb^k - y))$
3: **end for**

It is well known that the algorithm converges (in the sense that $f(b^k)$, where f is the objective functional, converges to the optimal value) under some conditions on the sequence of step sizes $\{t_k\}$. Valid choices include step sizes obeying $t_k < 2/\|X\|^2$ and step sizes obtained by backtracking line search, see Becker et al. (2011); Beck and Teboulle (2009). Further, one can use duality theory to derive concrete stopping criteria, see the supplementary appendix C (Bogdan et al., 2015) for details.

Many variants are of course possible and one may entertain accelerated proximal gradient methods in the spirit of FISTA, see Beck and Teboulle (2009) and Nesterov (2004, 2007). The scheme below is adapted from Beck and Teboulle (2009).

Algorithm 2 Accelerated proximal gradient algorithm for SLOPE equation (1.10)

Require: $b^0 \in \mathbb{R}^p$, and set $a^0 = b^0$ and $\theta_0 = 1$

1: **for** $k = 0, 1, \ldots$ **do**
2: $\quad b^{k+1} = \mathrm{prox}_{t_k J_\lambda}(a^k - t_k X'(Xa^k - y))$
3: $\quad \theta_{k+1}^{-1} = \frac{1}{2}(1 + \sqrt{1 + 4/\theta_k^2})$
4: $\quad a^{k+1} = b^{k+1} + \theta_{k+1}(\theta_k^{-1} - 1)(b^{k+1} - b^k)$
5: **end for**

The code in our numerical experiments uses a straightforward implementation of the standard FISTA algorithm, along with problem-specific stopping criteria. Standalone Matlab and R implementations of the algorithm are available at `http://www-stat.stanford.edu/~candes/SortedL1`. In addition, the TFOCS package available at `http://cvxr.com` Becker et al. (2011) implements Algorithms 1 and 2 as well as its many variants.

2.2 Fast prox algorithm

Given $y \in \mathbb{R}^p$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, the prox to the sorted ℓ_1 norm is the unique solution to

$$\mathrm{prox}(y; \lambda) := \mathrm{argmin}_{x \in \mathbb{R}^p} \ \tfrac{1}{2}\|y - x\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |x|_{(i)}. \tag{2.2}$$

A simple observation is this: at the solution to equation (2.2), the sign of each $x_i \neq 0$ will match that of y_i. It therefore suffices to solve the problem for $|y|$ and restore the signs in a post-processing step, if needed. Likewise, note that applying any permutation P to y results in a solution Px. We can thus choose a permutation that sorts the entries in y and apply its inverse to obtain the desired solution. Therefore, without loss of generality we can make the following assumption:

Assumption 2.1. *The vector y obeys $y_1 \geq y_2 \geq \cdots \geq y_p \geq 0$.*

The proposition below, proved in the supplementary appendix (Bogdan et al., 2015), provides a convenient reformulation of the proximal problem equation (2.2) by reformulating it as a quadratic program (QP).

Proposition 2.2. *Under Assumption 2.1 we can reformulate equation* (2.2) *as*

$$\begin{array}{ll} \textit{minimize} & \frac{1}{2}\|y-x\|_{\ell_2}^2 + \sum_{i=1}^p \lambda_i x_i \\ \textit{subject to} & x_1 \geq x_2 \geq \cdots \geq x_p \geq 0. \end{array} \tag{2.3}$$

We do not suggest performing the prox calculation by calling a standard QP solver applied to equation (2.3). Rather, we introduce the FastProxSL1 algorithm for computing the prox: for ease of exposition, we introduce Algorithm 3 in its simplest form before presenting a stack implementation (Algorithm 4) running in $O(p)$ flops, after an $O(p \log p)$ sorting step.

Algorithm 3, which terminates in at most p steps, is simple to understand: we simply keep on averaging until the monotonicity property holds, at which point the solution is known in closed form. The key point establishing the correctness of the algorithm is that the update does not change the value of the prox. This is formalized below.

Algorithm 3 FastProxSL1

input: Nonnegative and nonincreasing sequences y and λ.
while $y - \lambda$ is not nonincreasing **do**
Identify strictly increasing subsequences, i.e. segments $i : j$ such that

$$y_i - \lambda_i < y_{i+1} - \lambda_{i+1} < \cdots < y_j - \lambda_j. \tag{2.4}$$

Replace the values of y and λ over such segments by their average value: for $k \in \{i, i+1, \ldots, j\}$

$$y_k \leftarrow \frac{1}{j-i+1} \sum_{i \leq k \leq j} y_k, \qquad \lambda_k \leftarrow \frac{1}{j-i+1} \sum_{i \leq k \leq j} \lambda_k.$$

end while
output: $x = (y - \lambda)_+$.

Lemma 2.3. *The solution does not change after each update; formally, letting* (y^+, λ^+) *be the updated value of* (y, λ) *after one pass in Algorithm 3,*

$$\mathrm{prox}(y; \lambda) = \mathrm{prox}(y^+; \lambda^+).$$

Next, if $(y-\lambda)_+$ *is nonincreasing, then it is the solution to equation* (2.2)*, i.e.* $\mathrm{prox}(y; \lambda) = (y-\lambda)_+$.

This lemma, whose proof is in the supplementary appendix (Bogdan et al., 2015), guarantees that the FastProxSL1 algorithm finds the solution to equation (2.2) in a finite number of steps.

As stated earlier, it is possible to obtain a careful $O(p)$ implementation of FastProxSL1. Below we present a stack-based approach. We use tuple notation $(a, b)_i = (c, d)$ to denote $a_i = c$, $b_i = d$. For the complexity of the algorithm note that we create a total of p new tuples. Each of these tuples is merged into a previous tuple at most once. Since the merge takes a constant amount of time the algorithm has the desired $O(p)$ complexity.

With this paper, we are making available a C, a Matlab, and an R implementation of the stack-based algorithm at `http://www-stat.stanford.edu/~candes/SortedL1`. The algorithm is also implemented in R package SLOPE, available on CRAN, and included in the current version of the TFOCS package. Table 1 reports the average runtimes of the algorithm (MacBook Pro, 2.66GHz, Intel Core i7) when applied to vectors of fixed length and varying sparsity.

Algorithm 4 Stack-based algorithm for FastProxSL1.

input: Nonnegative and nonincreasing sequences y and λ.
 # Find optimal group levels
$t \leftarrow 0$
for $k = 1$ to n **do**
  $t \leftarrow t + 1$
  $(i, j, s, w)_t = (k,\ k,\ y_i - \lambda_i,\ (y_i - \lambda_i)_+)$
  while $(t > 1)$ and $(w_{t-1} \le w_t)$ **do**
    $(i, j, s, w)_{t-1} \leftarrow (i_{t-1},\ j_t,\ s_{t-1} + s_t, (\frac{j_{t-1} - i_{t-1} + 1}{j_t - i_{t-1} + 1} \cdot s_{t-1} + \frac{j_t - i_t + 1}{j_t - i_{i-1} + 1} \cdot s_t)_+)$
    Delete $(i, j, s, w)_t$, $t \leftarrow t - 1$
  end while
end for
 # Set entries in x for each block
for $\ell = 1$ to t **do**
  for $k = i_\ell$ to j_ℓ **do**
    $x_k \leftarrow w_\ell$
  end for
end for

	$p = 10^5$	$p = 10^6$	$p = 10^7$
Total prox time (sec.)	9.82e-03	1.11e-01	1.20e+00
Prox time after normalization (sec.)	6.57e-05	4.96e-05	5.21e-05

Table 1: Average runtimes of the stack-based prox implementation with normalization steps (sorting and sign changes) included, respectively excluded.

2.3 Related algorithms

Brad Efron informed us about the connection between the FastProxSL1 algorithm for SLOPE and a simple iterative algorithm for solving istonic problems called the pool adjacent violators algorithm (PAVA) Kruskal (1964); Barlow et al. (1972). A simple instance of an isotonic regression problem involves fitting data in a least squares sense in such a way that the fitted values are monotone:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|y - x\|_{\ell_2}^2 \\ \text{subject to} & x_1 \ge x_2 \ge \cdots \ge x_p \end{array} \tag{2.5}$$

here, y is a vector of observations and x is the vector of fitted values, which are here constrained to be nonincreasing. We have chosen this formulation to emphasize the connection with equation (2.3).

Indeed, our QP equation (2.3) is equivalent to

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\sum_{i=1}^{p}(y_i - \lambda_i - x_i)^2 \\ \text{subject to} & x_1 \geq x_2 \geq \cdots \geq x_p \geq 0 \end{array}$$

so that we see are really solving an isotonic regression problem with data $y_i - \lambda_i$. Algorithm 3 is then a version of PAVA as described in Barlow et al. (1972), see Best and Chakravarti (1990); Grotzinger and Witzgall (1984) for related work and connections with active set methods. Also, an elegant R package for isotone regression has been contributed by de Leeuw et al. (2009) and can be used to compute the prox to the sorted ℓ_1 norm.

Similar algorithms were also proposed in Zhong and Kwok (2012) to solve the OSCAR optimization problem defined as

$$\text{minimize} \quad \tfrac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda_1\|b\|_{\ell_1} + \lambda_2 \sum_{i<j} \max(|b_i|,\, |b_j|). \tag{2.6}$$

The OSCAR formulation was introduced in Bondell et al. (2008) to encourage grouping of correlated predictors. The OSCAR penalty term can expressed as $\sum_{i=1}^{p} \alpha_i |b|_{(i)}$ with $\alpha_i = \lambda_1 + (p-i)\lambda_2$; hence, this is a sorted ℓ_1 norm with a linearly decaying sequence of weights. Bondell et al. (2008) do not present a special algorithm for solving equation (2.6) other than casting the problem as a QP. In the article (Zeng and Figueiredo, 2014), which appeared after our manuscript was made publicly available, the OSCAR penalty term was further generalized to a Weigthed Sorted L-one norm, which coincides with the SLOPE formulation. This latter article does not discuss statistical properties of this fitting procedure.

3 Results

We now illustrate the performance of our SLOPE proposal in three different ways. First, we describe a multiple-testing situation where reducing the problem to a model selection setting and applying SLOPE assures FDR control, and results in a testing procedure with appreciable properties. Second, we discuss guiding principles to choose the sequence of λ_i's in general settings, and illustrate the efficacy of the proposals with simulations. Third, we apply SLOPE to a data set collected in genetics investigations.

3.1 An application to multiple testing

In this section we show how SLOPE can be used as an effective multiple comparison controlling procedure in a testing problem with a specific correlation structure. Consider the following situation. Scientists perform $p = 1,000$ experiments in each of 5 randomly selected laboratories, resulting in observations that can be modeled as

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \leq i \leq 1000,\ 1 \leq j \leq 5, \tag{3.1}$$

where the laboratory effects τ_j are i.i.d. $\mathcal{N}(0, \sigma_\tau^2)$ random variables and the errors $z_{i,j}$ are i.i.d. $\mathcal{N}(0, \sigma_z^2)$, with the τ and z sequences independent of each other. It is of interest to test whether $H_i : \mu_i = 0$ versus a two-sided alternative. Averaging the scores over all five labs, results in

$$\bar{y}_i = \mu_i + \bar{\tau} + \bar{z}_i, \quad 1 \leq i \leq 1000,$$

with $\bar{y} \sim \mathcal{N}(\mu, \Sigma)$ and $\Sigma_{i,i} = \frac{1}{5}(\sigma_\tau^2 + \sigma_z^2) = \sigma^2$ and $\Sigma_{i,j} = \frac{1}{5}\sigma_\tau^2 = \rho$ for $i \neq j$.

The problem has then been reduced to testing if the marginal means of a multivariate Gaussian vector with equicorrelated entries do not vanish. One possible approach is to use marginal tests based on $\bar{y}_i$'s and rely on the Benjamini-Hochberg procedure to control FDR. That is, we can order $|\bar{y}|_{(1)} \geq |\bar{y}|_{(2)} \geq \cdots \geq |\bar{y}|_{(p)}$ and apply the step-up procedure with critical values equal to $\sigma \cdot \Phi^{-1}(1 - iq/2p)$.

Another possible approach is to 'whiten the noise' and express our multiple testing problem in the form of a regression equation

$$\tilde{y} = \Sigma^{-1/2}\bar{y} = \Sigma^{-1/2}\mu + \epsilon, \tag{3.2}$$

where $\epsilon \sim \mathcal{N}(0, I_p)$. Treating $\Sigma^{-1/2}$ as the regression design matrix, our problem is equivalent to classical model selection: identify the non-zero components of the vector μ of regression coefficients.[5] Note that while the matrix Σ is far from being diagonal, $\Sigma^{-1/2}$ is diagonally dominant. For example when $\sigma^2 = 1$ and $\rho = 0.5$ then $\Sigma_{i,i}^{-1/2} = 1.4128$ and $\Sigma_{i,j}^{-1/2} = -0.0014$ for $i \neq j$. Thus, every low-dimensional sub-model obtained by selecting few columns of the design matrix $\Sigma^{-1/2}$ will be very close to orthogonal. In summary, the transformation (3.2) reduces the multiple-testing problem with strongly positively correlated test statistics to a problem of model selection under approximately orthogonal design, which is well suited for the application of SLOPE with the λ_{BH} values.

To compare the performances of these two approaches, we simulate data according to the model (3.1) with variance components $\sigma_\tau^2 = \sigma_z^2 = 2.5$, which yield $\sigma^2 = 1$ and $\rho = 0.5$. We consider a sequence of sparse settings, where the number k of nonzero μ_i's varies between 0 and 80. To obtain moderate power the nonzero means are set to $\sqrt{2\log p}/c \approx 2.63$, where c is the Euclidean norm of each of the columns of $\Sigma^{-1/2}$. We compare the performance of SLOPE and BH on marginal tests under two scenarios: (1) assuming $\sigma_\tau^2 = \sigma_z^2 = 2.5$ known, and (2) estimating them using the classical unweighted means method based on equating the ANOVA mean squares to their expectations:

$$\hat{\sigma}_z^2 = \text{MSE}, \quad \hat{\sigma}_\tau^2 = \frac{\text{MS}\tau - \text{MSE}}{1000};$$

using the standard notation from ANOVA analysis, MSE is the mean square due to the error in the model (3.1) and MSτ is the mean square due to the random factor τ. To use SLOPE, we center the vector $\tilde{y}$ by subtracting its mean, and center and standardize the columns of $\hat{\Sigma}^{-1/2}$, so they have zero means and unit l_2 norms. Figure 3 reports the results of these simulations, averaged over 500 independent replicates.

In our setting, the estimation procedure has no influence on SLOPE. Under both scenarios (variance components known and unknown) SLOPE keeps FDR at the nominal level as long as $k \leq 40$. Then its FDR slowly increases, but for $k \leq 80$ it is still very close to the nominal level as shown in Figure 3(c). In contrast, the performance of BH differs significantly: when σ^2 is known BH on the marginal tests is too conservative, with an average FDP below the nominal level, see Figure 3(a) and (b). When σ^2 is estimated, the average FDP of this procedure increases and for $q = 0.05$, it significantly exceeds the nominal level. Under both scenarios (known and unknown σ^2) the power of BH is substantially smaller than the power provided by SLOPE (Figure 3(d)). Moreover, the False Discovery Proportion (FDP) in the marginal tests with BH correction appears more variable across replicates than that of SLOPE (Figure 3a, 3b and 3c). Figure 4 presents the

[5] To be explicit, equation (3.2) is the basic regression model with $X = \Sigma^{-1/2}$ and $\beta = \mu$.

results in greater detail for $q = 0.1$ and $k = 50$: in approximately 65% of the cases the observed FDP for BH is equal to 0, while in the remaining 35% it takes values which are distributed over the whole interval (0,1). This behavior is undesirable. On the one hand, FDP = 0 typically equates with few discoveries (and hence power loss). On the other hand, if many FDP = 0 contribute to the average in the FDR, this quantity is kept below the desired level q even if, when there are discoveries, a large number of them are false. Indeed, in approximately 26% of all cases BH on the marginal tests did not make any rejections (i.e., $R = 0$); and conditional on $R > 0$, the mean of FDP is equal to 0.16 with a standard deviation of 0.28, which clearly shows that the observed FDP is typically far away from the nominal value of $q = 0.1$. In other words, while BH is close to controlling the FDR, the scientists would either make no discoveries or have very little confidence on those actually made. In contrast, SLOPE results in a more predictable FDP and a substantially larger and more predictable True Positive Proportion (TPP, fraction of correctly identified true signals), see Figure 4.

3.2 Choosing λ in general settings.

In the previous sections we observed that, for orthogonal designs, Lasso with $\lambda_{\text{Bonf}} = \sigma \cdot \Phi^{-1}(1 - \alpha/2p)$ controls FWER at the level α, while SLOPE with the sequence $\lambda = \lambda_{\text{BH}}$ controls FDR at the level q. We are interested, however, in applying these procedures in more general settings, specifically when $p > n$ and there is some correlation among the explanatory variables, and when the value of σ^2 is not known. We start tackling the first situation. Correlation among regressors notoriously introduces a series of complications in the statistical analysis of linear models, ranging from the increased computational costs that motivated the early popularity of orthogonal designs, to the conceptual difficulties of distinguishing causal variables among correlated ones. Indeed, recent results on the consistency of ℓ_1 penalization methods typically require some form of partial orthogonality. SLOPE and Lasso aim at finite sample properties, but it would not be surprising if departures from orthogonality were to have a serious effect. To explore this, we study the performance of Lasso and SLOPE in the case where the entries of the design matrix are generated independently from the $\mathcal{N}(0, 1/n)$ distribution. Specifically, we consider two Gaussian designs with $n = 5000$: one with $p = 2n$ =10,000 and one with $p = n/2 = 2500$. We set the value of non-zero coefficients to $5\sqrt{2 \log p}$ and consider situations where the number of important variables ranges between 0 and 100. Figure 5 illustrates that under such Gaussian designs both Lasso-λ_{Bonf} and SLOPE loose the control over their targeted error rates (FWER and FDR) as the number k of nonzero coefficients increases, with a departure that is more severe when the ratio between p/n is larger.

3.2.1 The effect of shrinkage

What is behind this fairly strong effect, and is it possible to choose a λ sequence to compensate it? Some useful insights come from studying the solution of the Lasso. Assume that the columns of X have unit norm and that $z \sim \mathcal{N}(0, 1)$. Then the optimality conditions for the Lasso give

$$\begin{aligned} \hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) &= \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - X\beta - z)) \\ &= \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z), \end{aligned} \tag{3.3}$$

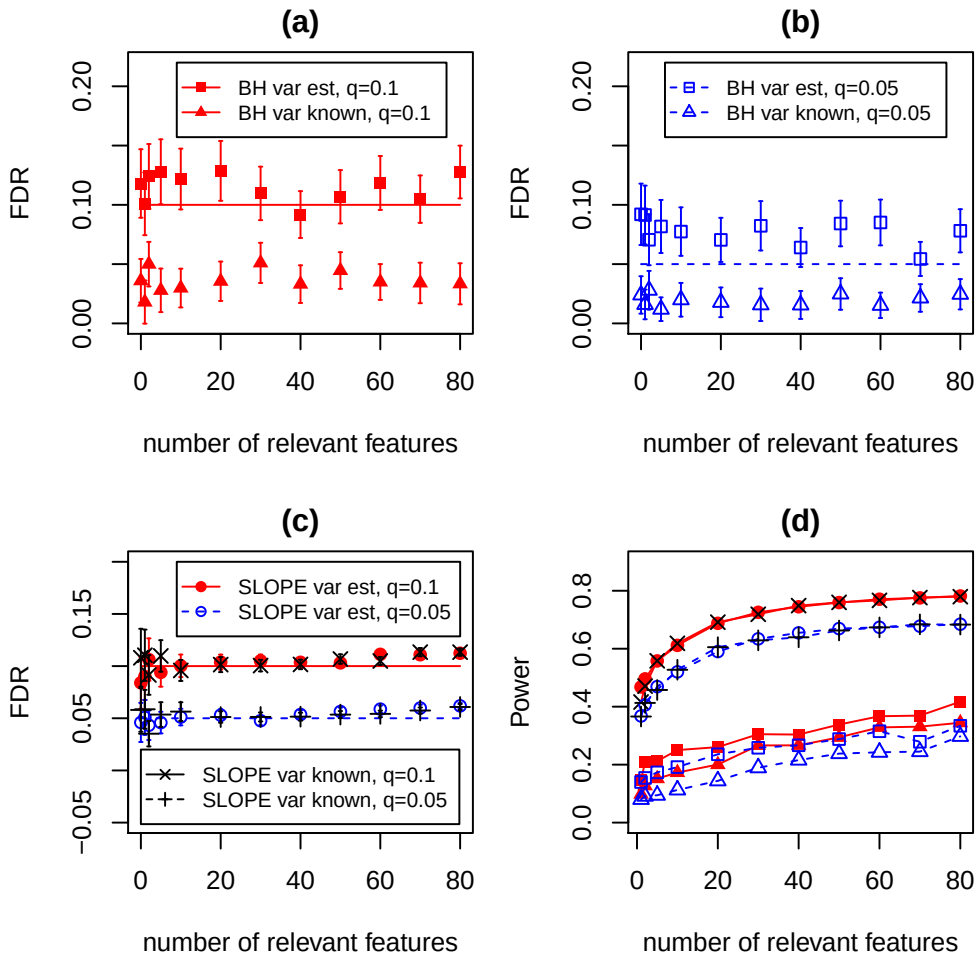


Figure 3: Simulation results for testing multiple means from correlated statistics. (a)–(b) Mean FDP $\pm$ 2 SE for marginal tests as a function of k. (c) Mean FDP $\pm$ 2 SE for SLOPE. (d) Power plot.

where η_λ is the soft-thresholding operator, $\eta_\lambda(t) = \mathrm{sgn}(t)(|t|-\lambda)_+$, applied componentwise. Defining $v_i = \langle X_i, \sum_{j\neq i} X_j(\beta_j - \hat{\beta}_j)\rangle$, we can write

$$\hat{\beta}_i = \eta_\lambda(\beta_i + X_i'z + v_i), \tag{3.4}$$

which expresses the relation between the estimated value of $\hat{\beta}_i$ and its true value β_i. If the variables are orthogonal, the v_i's are identically equal to 0, leading to $\hat{\beta}_i = \eta_\lambda(\beta_i + X_i'z)$. Conditionally on X, $X_i'z \sim \mathcal{N}(0,1)$ and by using Bonferroni's method, one can choose λ such that $\mathbb{P}(\max_i |X_i'z| > \lambda) \leq \alpha$. When X is not orthogonal, however, $v_i \neq 0$ and its size increases with the estimation error of β_j (for $i \neq j$)—which depends on the magnitude of the shrinkage parameter λ. Therefore, even in the perfect situation where all the k relevant variables, and those alone, have been selected, and when all columns of the design matrix are realizations of independent random variables v_i will not be zero. Rather, the squared magnitude v_i^2 will be on the order of $\lambda^2 \cdot k/n$. In other words, the

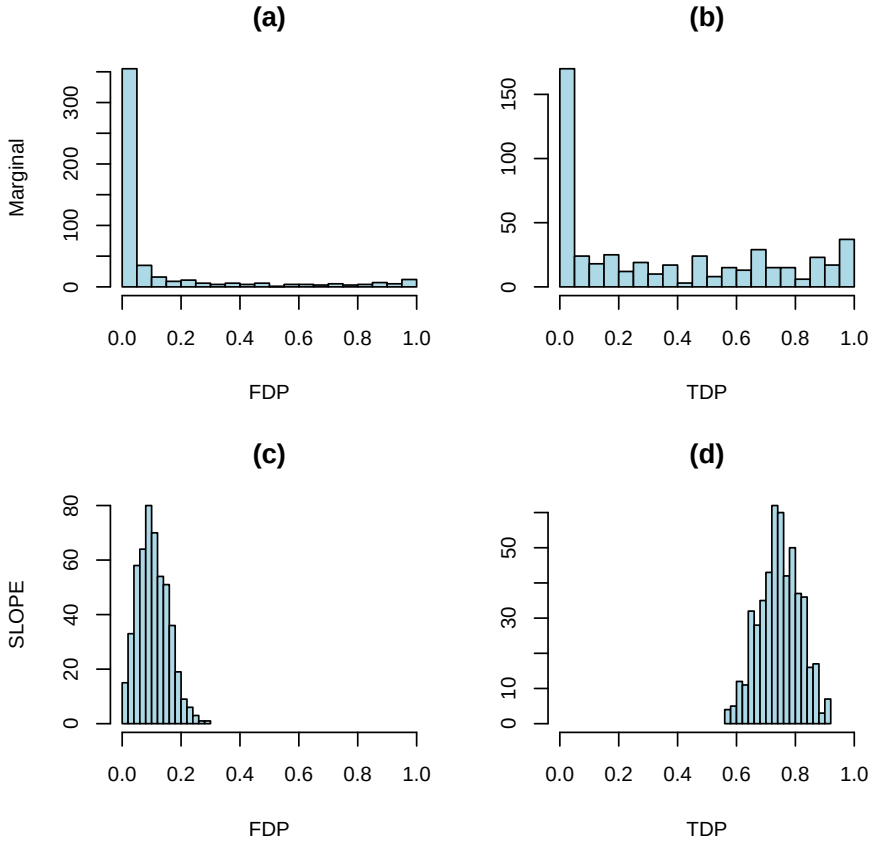
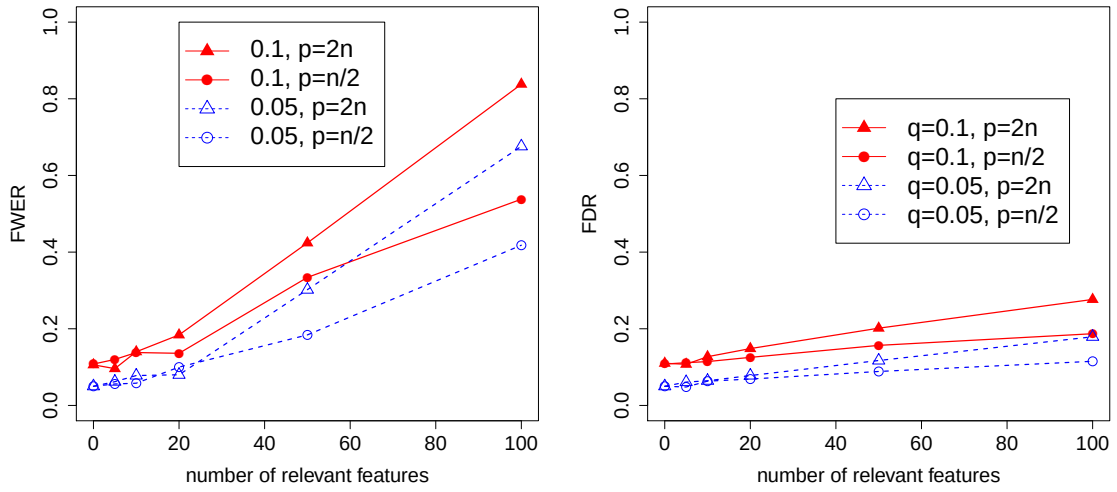


Figure 4: Testing example with $q = 0.1$ and $k = 50$. The top row refers to marginal tests, and the bottom row to SLOPE. Both procedures use the estimated variance components. Histograms of false discovery proportions are in the first column and of true positive proportions in the second.

variance that would determine the correct Bonferroni threshold is on the order $1 + \lambda^2 \cdot k/n$. In reality, the true k is not known a priori, and the selected k depends on the value of the smoothing parameter λ, so that it is not trivial to implement this correction in the Lasso. SLOPE, however, uses a decreasing sequence λ, analogous to a step-down procedure, and this extra noise due to the shrinkage of relevant variables can be incorporated by progressively modifying the λ sequence. In evocative, if not exact terms, λ_1 is used to select the first variable to enter the model: at this stage we are not aware of any variable whose shrunk coefficient is 'effectively increasing' the noise level, and we can keep $\lambda_1 = \lambda_{\mathrm{BH}}(1)$. The value of λ_2 determines the second variable to enter the model and, hence, we know that there is already one important variable whose coefficient has been shrunk by roughly $\lambda_{\mathrm{BH}}(1)$: we can use this information to re-define λ_2. Similarly, when using λ_3 to identify the third variable, we know of two relevant regressors whose coefficients have been shrunk by amounts determined by λ_1 and λ_2, and so on. What follows is an attempt to make this intuition more precise, accounting for the fact that the sequence λ needs to be determined a priori, and we need to make a prediction on the values of the cross products $X_i' X_j$ appearing in the definition of v_i. Before we turn to this, we want to underscore how this explanation for the loss



(a) FWER of Lasso with λ_{Bonf}.

(b) FDR of SLOPE with λ_{BH}.

Figure 5: Observed FWER for Lasso with λ_{Bonf} and FDR for SLOPE with λ_{BH} under Gaussian design and $n = 5,000$. The results are averaged over 500 replicates.

of FDR control is consistent with patterns evident from Figure 5: the problem is more serious as k increases (and hence the effect of shrinkage is felt on a larger number of variables), and as the ratio p/n increases (which for Gaussian designs results in larger empirical correlation $|X_i' X_j|$). Our loose analysis suggests that when k is really small, SLOPE with λ_{BH} yields an FDR that is close to the nominal level, as empirically observed.

3.2.2 Adjusting the regularizing sequence for SLOPE

In light of equation (3.4) we would like an expression for $X_i' X_{\mathcal{S}}(\beta_{\mathcal{S}} - \hat{\beta}_{\mathcal{S}})$, where with $\mathcal{S}$, $X_{\mathcal{S}}$ and $\beta_{\mathcal{S}}$ we indicate the support of β, the subset of variables associated to $\beta_i \neq 0$, and the value of their coefficients, respectively.

Again, to obtain a very rough evaluation of the SLOPE solution, we can start from the Lasso. Let us assume that the size of $\beta_{\mathcal{S}}$ and the value of λ are such that the support and the signs of the regression coefficients are correctly recovered in the solution. That is, we assume that $\text{sign}(\beta_j) = \text{sign}(\hat{\beta}_j)$ for all j, with the convention that $\text{sign}(0) = 0$. Without loss of generality, we further assume that $\beta_j \geq 0$. Now, the Karush–Kuhn-Tucker (KKT) optimality conditions for Lasso yield

$$X_S'(y - X\hat{\beta}_S) = \lambda \cdot 1_S, \tag{3.5}$$

implying

$$\hat{\beta}_S = (X_S' X_S)^{-1}(X_S' y - \lambda \cdot 1_S).$$

In the case of SLOPE, rather than one λ, we have a sequence $\lambda_1, \ldots, \lambda_p$. Assuming again that this is chosen so that we recover exactly the support $\mathcal{S}$, the estimates of the nonzero components are very roughly equal to

$$\hat{\beta}_S = (X_{\mathcal{S}}' X_{\mathcal{S}})^{-1}(X_{\mathcal{S}}' y - \lambda_{\mathcal{S}}) = \hat{\beta}_{OLS} - (X_{\mathcal{S}}' X_{\mathcal{S}})^{-1}\lambda_{\mathcal{S}},$$

where $\lambda_{\mathcal{S}} = (\lambda_1, \ldots, \lambda_{|S|})'$ and $\hat{\beta}_{\mathrm{OLS}}$ is the least-squares estimator of β_S. This leads to $\mathbb{E}(\beta_{\mathcal{S}} - \hat{\beta}_S) \approx (X'_{\mathcal{S}} X_{\mathcal{S}})^{-1} \lambda_{\mathcal{S}}$ and

$$\mathbb{E}\, X_i' X_{\mathcal{S}} (\beta_{\mathcal{S}} - \hat{\beta}_{\mathcal{S}}) \approx \mathbb{E}\, X_i' X_{\mathcal{S}} (X'_{\mathcal{S}} X_{\mathcal{S}})^{-1} \lambda_{\mathcal{S}},$$

an expression that tells us the typical size of v_i in equation (3.4).

For the case of Gaussian designs, where the entries of X are i.i.d. $\mathcal{N}(0, 1/n)$, for $i \notin \mathcal{S}$,

$$\mathbb{E}(X_i' X_{\mathcal{S}} (X'_{\mathcal{S}} X_{\mathcal{S}})^{-1} \lambda_{\mathcal{S}})^2 = \frac{1}{n} \lambda'_{\mathcal{S}}\, \mathbb{E}(X'_{\mathcal{S}} X_{\mathcal{S}})^{-1} \lambda_{\mathcal{S}} = w(|\mathcal{S}|) \cdot \|\lambda_{\mathcal{S}}\|_{\ell_2}^2, \quad w(k) = \frac{1}{n-k-1}. \tag{3.6}$$

This uses the fact that the expected value of an inverse $k \times k$ Wishart with n degrees of freedom is equal to $I_k/(n-k-1)$.

This suggests the sequence of λ's described below denoted by λ_{G} since it is motivated by Gaussian designs. We start with $\lambda_{\mathrm{G}}(1) = \lambda_{\mathrm{BH}}(1)$. At the next stage, however, we need to account for the slight increase in variance so that we do not want to use $\lambda_{\mathrm{BH}}(2)$ but rather

$$\lambda_{\mathrm{G}}(2) = \lambda_{\mathrm{BH}}(2) \sqrt{1 + w(1) \lambda_{\mathrm{G}}(1)^2}.$$

Continuing, this gives

$$\lambda_{\mathrm{G}}(i) = \lambda_{\mathrm{BH}}(i) \sqrt{1 + w(i-1) \sum_{j<i} \lambda_{\mathrm{G}}(j)^2}. \tag{3.7}$$

Figure 6 plots the adjusted values given by equation (3.7). As is clear, these new values yield a procedure that is more conservative than that based on λ_{BH}. It can be observed that the corrected

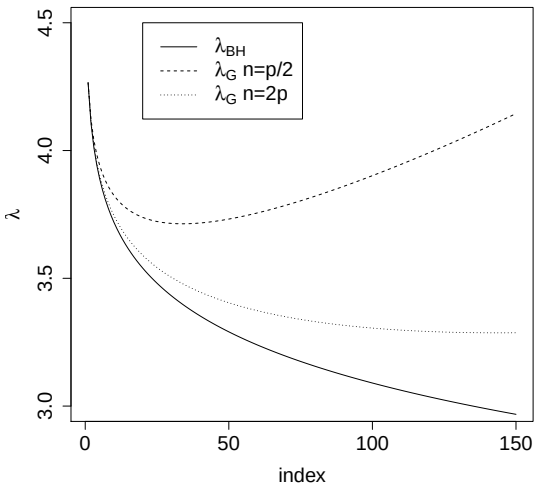
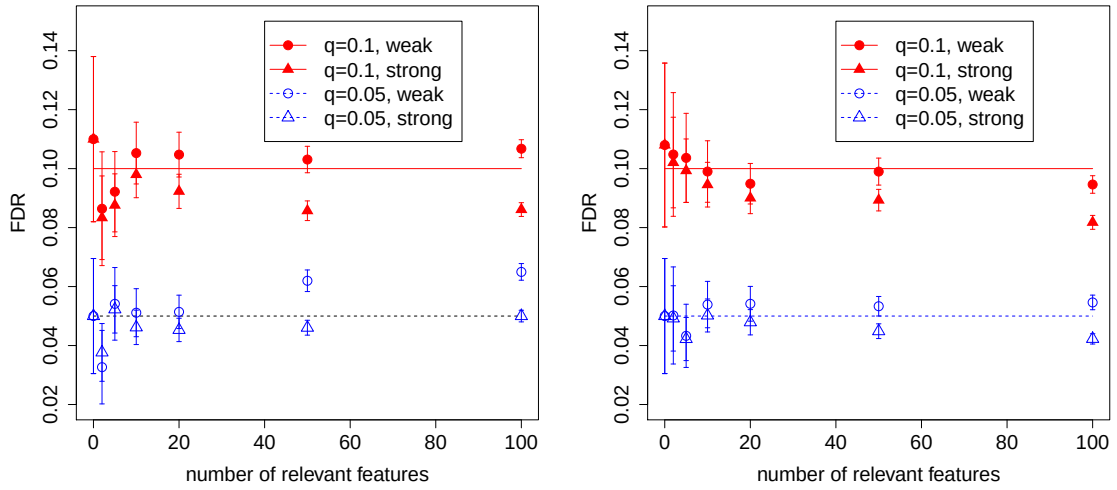


Figure 6: Graphical representation of sequences $\{\lambda_i\}$ for $p = 5000$ and $q = 0.1$. The solid line is λ_{BH}, the dashed (resp. dotted) line is λ_{G} given by equation (3.7) for $n = p/2$ (resp. $n = 2p$).

sequence $\lambda_{\mathrm{G}}(i)$ may no longer be decreasing (as in the case where $n = p/2$ in the figure). It would not make sense to use such a sequence—note that SLOPE would no longer be convex—and letting $k^\star = k(n, p, q)$ be the location of the global minimum, we shall work with

$$\lambda_{\mathrm{G}^\star}(i) = \begin{cases} \lambda_{\mathrm{G}}(i), & i \le k^\star, \\ \lambda_{k^\star}, & i > k^\star, \end{cases} \quad \text{with } \lambda_{\mathrm{G}}(i) \text{ as in equation (3.7).} \tag{3.8}$$

An immediate validation—if the intuition that we have stretched this far has any bearing in reality—is the performance of $\lambda_{\mathrm{G}^\star}$ in the setup of Figure 5. In Figure 7 we illustrate the performance of SLOPE for large signals $\beta_i = 5\sqrt{2\log p}$ as in Figure 5, as well as for rather weak signals with $\beta_i = \sqrt{2\log p}$. The correction works very well, rectifying the loss of FDR control documented in Figure 5. For $p = 2n = 10,000$, the values of the critical point $k^\star$ are 51 for $q = 0.05$ and 68 for $q = 0.1$. For $p = n/2 = 2,500$, they become 95 and 147 respectively. It can be observed that for large signals, SLOPE keeps FDR below the nominal level even after passing the critical point. Interestingly, the control of FDR is more difficult when the coefficients have small amplitudes. We believe that some increase of FDR for weak signals is related to the loss of power, which our correction does not account for. However, even for weak signals the observed FDR of SLOPE with $\lambda_{\mathrm{G}^\star}$ is very close to the nominal level when $k \le k^\star$.



(a) $p = 2n = 10,000$. **(b)** $p = n/2 = 2,500$.

Figure 7: Mean FDP ± 2SE for SLOPE with $\lambda_{\mathrm{G}^\star}$. Strong signals have nonzero regression coefficients set to $5\sqrt{2\log p}$ while this value is set to $\sqrt{2\log p}$ for weak signals.

In situations where one cannot assume that the design is Gaussian or that columns are independent, we suggest replacing $w(i-1)\sum_{j<i}\lambda_j^2$ in the formula (3.7) with a Monte Carlo estimate of the correction. Let X denote the standardized version of the design matrix, so that each column has a mean equal to zero and unit l_2 norm. Suppose we have computed $\lambda_1, \ldots, \lambda_{i-1}$ and wish to compute λ_i. Let $X_{\mathcal{S}}$ indicate a matrix formed by selecting those columns with indices in some set $\mathcal{S}$ of cardinality $i-1$ and let $j \notin \mathcal{S}$. After randomly selecting $\mathcal{S}$ and j, the correction can be approximated by the average of $(X_j' X_{\mathcal{S}} (X_{\mathcal{S}}' X_{\mathcal{S}})^{-1} \lambda_{1:i-1})^2$ across realizations, where $\lambda_{1:i-1} = (\lambda_1, \ldots, \lambda_{i-1})'$.

Significantly more research is needed to understand the properties of this heuristics and to design more efficient alternatives. Our simulations so far suggest that it provides approximate FDR control when looking at the average across all possible signal placements, and—for any fixed signal location—if the columns of the design matrix are exchangeable. It is important to note that the computational cost of this procedure is relatively low. Two elements contribute to this. First, the complexity of the procedure is reduced by the fact that the sequence of λ's does not need to

be estimated entirely, but only up to the point $k^\star$ where it start increasing (or simply flattens) and only for a number of entries on the order of the expected number of nonzero coefficients. Second, the smoothness of λ assures that it is enough to estimate λ on a grid of points between 1 and $k^\star$, making the problem tractable also for very large p. In Bogdan et al. (2013) we applied a similar procedure for the estimation of the regularizing sequence with $p = 2048^2 = 4194304$ and $n = p/5$ and found out that it was sufficient to estimate this sequence at only 40 grid points.

3.2.3 Unknown σ

According to formulas (1.5) and (1.10) the penalty in SLOPE depends on the standard deviation σ of the error term. In many applications σ is not known and needs to be estimated. When n is larger than p, this can easily be done by means of classical unbiased estimators. When $p \geq n$, some solutions for simultaneous estimation of σ and regression coefficients using ℓ_1 optimization schemes were proposed, see e.g., Städler et al. (2010) and Sun and Zhang (2012). Specifically, Sun and Zhang (2012) introduced a simple iterative version of the Lasso called the *scaled Lasso*. The idea of this algorithm can be applied to SLOPE, with some modifications. For one, our simulation results show that, under very sparse scenarios, it is better to de-bias the estimates of regression parameters by using classical least squares estimates within the selected model to obtain an estimate of σ^2.

We present our algorithm below. There, λ^S is the sequence of SLOPE parameters designed to work with $\sigma = 1$, obtained using the methods from Section 3.2.2.

Algorithm 5 Iterative SLOPE fitting when σ is unknown

input: y, X and initial sequence λ^S (computed for $\sigma = 1$)
initialize: $S_+ = \emptyset$
repeat
  $S = S_+$
  compute the RSS obtained by regressing y onto variables in S
  set $\hat{\sigma}^2 = \mathrm{RSS}/(n - |S| - 1)$
  compute the solution $\hat{\beta}$ to SLOPE with parameter sequence $\hat{\sigma} \cdot \lambda^S$
  set $S_+ = \mathrm{supp}(\hat{\beta})$
until $S_+ = S$

The procedure starts by using a conservative estimate of the standard deviation of the error term $\hat{\sigma}^{(0)} = \mathrm{Std}(y)$ and a related conservative version of SLOPE with $\lambda^{(0)} = \hat{\sigma}^{(0)} \cdot \lambda^S$. Then, in consecutive runs $\hat{\sigma}^{(k)}$ is computed using residuals from the regression model, which includes variables identified by SLOPE with sequence $\sigma^{(k-1)} \cdot \lambda^S$. The procedure is repeated until convergence, i.e., until the next iteration results in exactly the same model as the current one.

3.2.4 Simulations with idealized GWAS data

We illustrate the performance of the 'scaled' version of SLOPE and of our algorithm for the estimation of the parameters λ_i with simulations designed to mimic an idealized version of Genome Wide Association Studies (GWAS). We set $n = p = 5000$, and simulate 5000 genotypes of p independent Single Nucleotide Polymorphisms (SNPs). For each of these SNPs the minor allele frequency (MAF) is sampled from the uniform distribution on the interval $(0.1, 0.5)$. Let us underscore that this assumption of independence is not met in actual GWAS, where the number of typed SNPs is in

the order of millions. Rather, one can consider our data generating mechanism as an approximation of the result of preliminary screening of genotype variants to avoid complications due to correlation. Our goal here is not to argue that SLOPE has superior performance in GWAS, but rather to illustrate the computational costs and inferential results of our algorithms. The explanatory variables are defined as

$$\tilde{x}_{ij} = \begin{cases} -1 & \text{for} \quad aa \\ 0 & \text{for} \quad aA \\ 1 & \text{for} \quad AA \end{cases}, \tag{3.9}$$

where a and A denote the minor and reference alleles at the jth SNP for the ith individual. Then the matrix $\tilde{X}$ is centered and standardized, so the columns of the final design matrix X have zero mean and unit norm. The trait values are simulated according to the model

$$y = X\beta + z, \tag{3.10}$$

where $z \sim N(0, I)$, that is we assume only additive effects and no interaction between loci (epistasis). We vary the number of nonzero regression coefficients k between 0 and 50 and we set their size to $1.2\sqrt{2\log p} \approx 4.95$ ('moderate' signal). For for each value of k, 500 replicates are performed, in each selecting randomly among the columns of X, the k with non zero coefficients. Since our design matrix is centered and does not contain an intercept, we also center the vector of responses and let SLOPE work with $\tilde{y} = y - \bar{y}$, where $\bar{y}$ is the mean of y.

We set $q = 0.05$ and estimate the sequence λ via the Monte Carlo approach described in Section 3.2.2; here, we use 5,000 independent random draws of $X_{\mathcal{S}}$ and X_j to compute the next term in the sequence. The calculations terminated in about 90 seconds (HP EliteDesk 800 G1 TWR, 3,40GHz, Intel i7-4770) at λ_{31}, where the estimated sequence λ obtained a first local minimum. Figure 8(a) illustrates that up to this first minimum the Monte Carlo sequence λ_{MC} coincides with the heuristic sequence $\lambda_{\mathrm{G}^\star}$ for Gaussian matrices. In the result the FDR and power of 'scaled' SLOPE are almost the same for both sequences (Figures 8(b) and 8(c)).

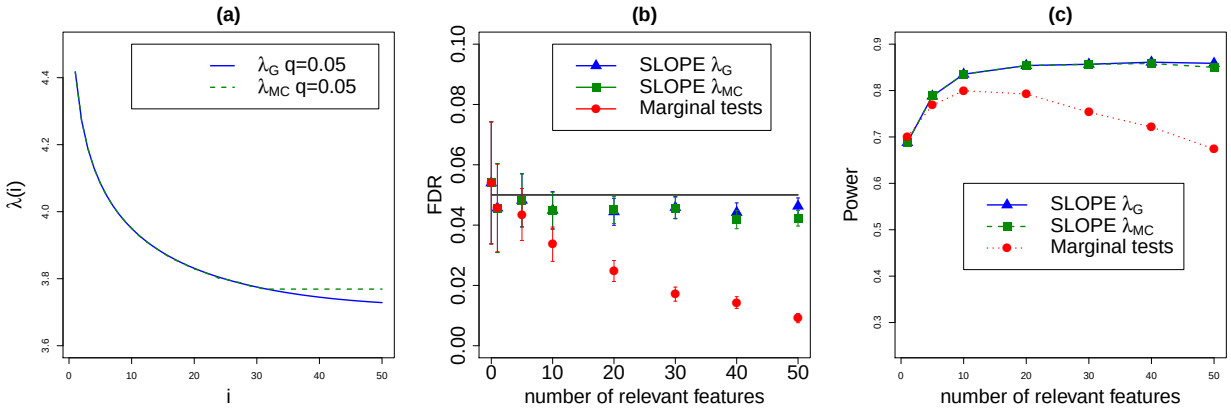


Figure 8: (a) Graphical representation of sequences λ_{MC} and λ_{G} for the SNP design matrix. (b) Mean FDP $\pm$ 2SE for SLOPE with $\lambda_{\mathrm{G}^\star}$ and λ_{MC} and for BH as applied to marginal tests. (c) Power of both versions of SLOPE and BH on marginal tests for $\beta_1 = \cdots = \beta_k = 1.2\sqrt{2\log p} \approx 4.95$, $\sigma = 1$. In each replicate, the signals are randomly placed over the columns of the design matrix, and the plotted data points are averages over 500 replicates.

In our simulations, the proposed algorithm for scaled SLOPE converges very quickly. The conservative initial estimate of σ leads to a relatively small model with few false discoveries since $\sigma^{(0)} \cdot \lambda^S$ controls the FDR in sparse settings. Typically, iterations to convergence see the estimated value of σ decrease and the number of selected variable increase. Since some signals remain undetected (the power is usually below 100%), σ is slightly overestimated at the point of convergence, which translates into controlling the FDR at a level slightly below the nominal one, see Figure 8(b).

Figures 8(b) and (c) compare scaled SLOPE with the "marginal" tests. The latter are based on t-test statistics

$$t_i = \hat{\beta}_i / \hat{\sigma}^2, \qquad \hat{\sigma}^2 = \mathrm{RSS}_i/(n-2),$$

where $\hat{\beta}_i$ (resp. RSS_i) is the least-square estimate of the regression coefficient (resp. the residual sum of squares) in the simple linear regression model including only the ith SNP. To adjust for multiplicity, we use BH at the nominal FDR level $q = 0.05$.

It can be observed that SLOPE and marginal tests do not differ substantially when $k \leq 5$. However, for $k \geq 10$ the FDR of the marginal tests approach falls below the nominal level and the power decreases from 80% for $k = 10$ to 67% for $k = 50$. SLOPE's power remains, instead, stable at the level of approximately 86% for $k \in \{20, \ldots, 50\}$. This conservative behavior of marginal tests results from the inflation of the noise level estimate caused by regressors that are unaccounted for in the simple regression model.

We use this idealized GWAS setting to explore also the effect of some model miss-specification. Firstly, we consider a trait y on which genotypes have effects that are not simply additive. We formalize this via the matrix $\tilde{Z}$ collecting the "dominant" effects

$$\tilde{z}_{ij} = \begin{cases} -1, & \text{for } aa, AA, \\ 1, & \text{for } aA. \end{cases} \tag{3.11}$$

The final design matrix $[X, Z]$ has the columns $[\tilde{X}, \tilde{Z}]$ centered and standardized. Now the trait values are simulated according to the model

$$y = [X, Z][\beta_X', \beta_Z']' + \epsilon,$$

where $\epsilon \sim N(0, I)$, the number of 'causal' SNPs k varies between 0 and 50, each causal SNP has an additive effect (non-zero components of β_X) equal to $1.2\sqrt{2\log p} \approx 4.95$ and a dominant effect (non-zero components of β_Z) randomly sampled from $N(0, \sigma = 2\sqrt{2\log p})$. The data is analyzed using model (3.10); that is, assuming linear effect of alleles even when this is not true.

Secondly, to explore the sensitivity to violations of the assumption of the normality of the error terms we considered (1) error terms z_i with a Laplace distribution and a scale parameter adjusted to that the variance is equal to one, and (2) error terms contaminated with 50 outliers $\sim N(0, \sigma = 5)$ representing 1% of all observations.

Figure 9 summarizes the performance of SLOPE and of the marginal tests (adjusted for multiplicity via BH), which we include for reference purposes. Violation of model assumption appear to affect power rather than FDR in the case of SLOPE. Specifically, in all three examples FDR is kept very close to the nominal level while the power is somewhat diminished with respect to Figure 8. The smallest difference is observed in case of Laplace errors, where the results of SLOPE are almost the same as in case of normal errors. This is also the case where the difference in performance due to model misspecification is negligible for marginal tests. In all other cases, this approach seems to be much more sensitive than SLOPE to model misspecification.

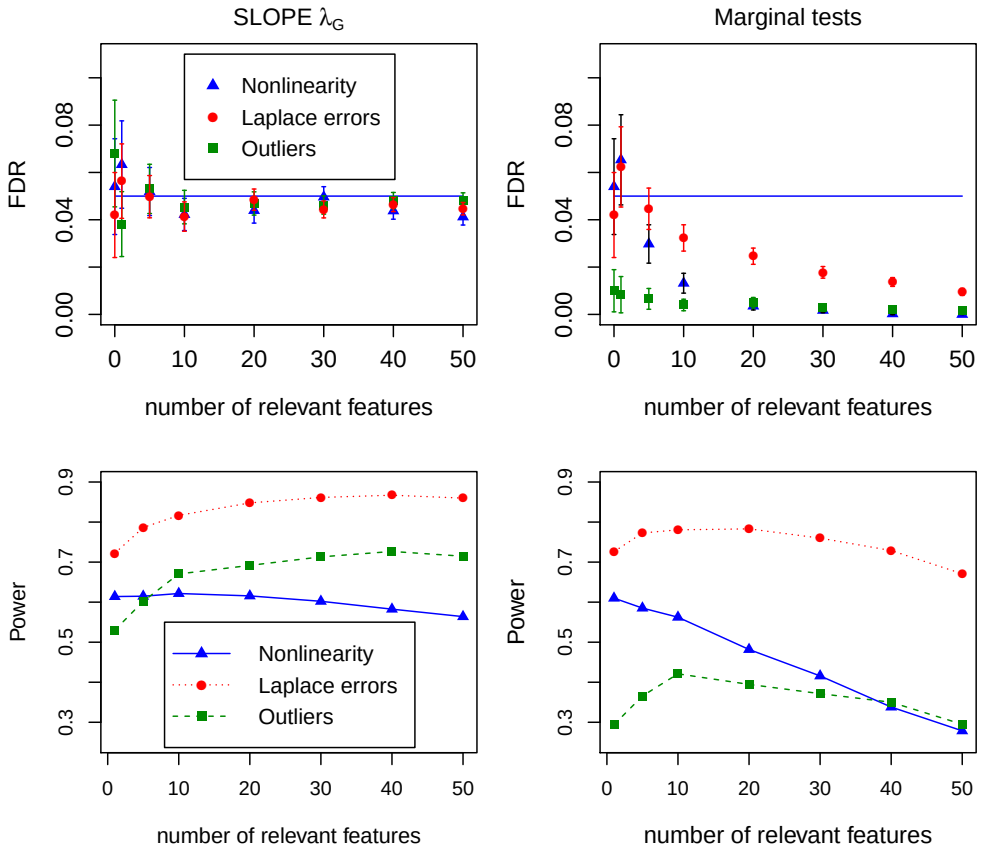


Figure 9: FDR and power of 'scaled' SLOPE based on "gaussian" sequence $\lambda_{G^\star}$ (left panel) and BH-corrected single marker tests (right panel) for different deviations from the assumed regression model. Error bars for FDR correspond to mean FDP $\pm$ 2 SE.

3.3 A real data example from genetics

In this section we illustrate the application of SLOPE to a current problem in genetics. In Service et al. (2014), the authors investigate the role of genetic variants in 17 regions in the genome, selected on the basis of previously reported association with traits related to cardiovascular health. Polymorphisms are identified via exome re-sequencing in approximately 6,000 individuals of Finnish descent: this provides a comprehensive survey of the genetic diversity in the coding portions of these regions and affords the opportunity to investigate which of these variants have an effect on the traits of interest. While the original study has a broader scope, we here tackle the problem of identifying which genetic variants in these regions impact the fasting blood HDL levels. Previous literature reported associations between 9 of the 17 regions and HDL, but the resolution of these earlier studies was unable to pinpoint to specific variants in these regions or to distinguish if only one or multiple variants within the regions impact HDL. The resequencing study was designed to address this problem.

The analysis in Service et al. (2014) relies substantially on "marginal" tests: the effect of each variant on HDL is examined via a linear regression that has cholesterol level as outcome and the genotype of the variant as explanatory variable, together with covariates that capture

possible population stratification. Such marginal tests are common in genetics and represent the standard approach in genome-wide association studies (GWAS). Among their advantages it is worth mentioning that they allow to use all available observations for each variant without requiring imputation of missing data; their computational cost is minimal; and they result in a p-value for each variant that can be used to clearly communicate to the scientific community the strength of the evidence in favor of its impact on a particular trait. Marginal tests, however, cannot distinguish if the association between a variant and a phenotype is "direct" or due to correlation between the variant in question and another, truly linked to the phenotype. Since most of the correlation between genetic variants is due to their location along the genome (with nearby variants often correlated), this confounding is often considered not too serious a limitation in GWAS: multiple polymorphisms associated to a phenotype in one locus simply indicate that there is at least one genetic variant (most likely not measured in the study) with impact on the phenotype in the locus. The situation is quite different in the re-sequencing study we want to analyze, where establishing if one or more variants in the same region influence HDL is one of the goals. To address this, the authors of Service et al. (2014) resort to regressions that include two variables at the time: one of these being the variant with previously documented strongest marginal signal in the region, the other variants that passed an FDR controlling threshold in the single variant analysis. Model selection strategies were only cursory explored with a step-wise search routine that targets BIC. Such limited foray into model selection is motivated by the fact that one major concern in genetics is to control some global measure of type I error and currently available model selection strategies do not offer finite sample guarantees with this regard. This goal is in line with that of SLOPE and so it is interesting for us to apply this new procedure to this problem.

The dataset in Service et al. (2014) comprises 1,878 variants, on 6,121 subjects. Before analyzing it with SLOPE, or other model selection tools, we performed the following filtering. We eliminated from considerations variants observed only once (a total of 486), since it would not be possible to make inference on their effect without strong assumptions. We examined correlation between variants and selected for analysis a set of variants with pair-wise correlation smaller than 0.3: larger values would make it quite challenging to interpret the outcomes; they render difficult the comparison of results across procedures since these might select different variables from a group of correlated ones; and large correlations are likely to adversely impact the efficacy of any model selection procedure. This reduction was carried out in an iterative fashion, selecting representative from groups of correlated variables, starting from stronger levels of correlation and moving onto lower ones. Among correlated variables, we selected those that had stronger univariate association with HDL, larger minor allele frequency (diversity), and, among very rare variants we privileged those whose annotation was more indicative of possible functional effects. Once variables were identified, we eliminated subjects that were missing values for more than 10 variants, and for HDL. The remaining missing values were imputed using the average allele count per variant. This resulted in a design with 5,375 subjects and 777 variants. The minor allele frequency of the variants included ranges from 2×10^{-4} to 0.5, with a median of 0.001 and a mean of 0.028: the data set still includes a number of rare variants, with the minor allele frequency smaller than 0.01.

In Service et al. (2014), association between HDL and polymorphisms was analyzed only for variants in regions previously identified as having an influence on HDL: *ABCA1, APOA1, CEPT, FADS1, GALNT2, LIPC, LPL, MADD*, and *MVK* (regions are identified with the name of one of the genes they contain). Moreover, only variants with minor allele frequencies larger than 0.01 were individually investigated, while non synonimous rare variants were analyzed with "burden

tests." These restrictions were motivated, at least in part, by the desire to reduce tests to the most well powered ones, so that controlling for multiple comparisons would not translate in an excessive decrease of power. Our analysis is based on all variants that survive the described filtering in all regions, including those not directly sequenced in the experiment in Service et al. (2014), but included in the study as landmarks of previously documented associations (*array SNPs* in the terminology of the paper). We compare the following approaches: the (1) marginal tests described above in conjunction with BH and $q = 0.05$; (2) BH and $q = 0.05$ applied to the p-values from the full model regression; (3) Lasso with λ_{Bonf} and $\alpha = 0.05$; (4) Lasso with λ_{CV} (in these last two cases we use the routines implemented in `glmnet` in R); (5) the R routine Step.AIC in forward direction and BIC as optimality criteria; (6) the R routine Step.AIC in backwards direction and BIC as optimality criteria; (7) SLOPE with $\lambda_{\text{G}^\star}$ and $q = 0.05$; (8) SLOPE with λ obtained via Monte Carlo starting from our design matrix. Defining the λ for Lasso–λ_{Bonf} and SLOPE requires a knowledge of the noise level σ^2: we estimated this from the residuals of the full model. When estimating λ via the Monte Carlo approach, for each i we used 5,000 independent random draws of $X_{\mathcal{S}}$ and X_j. Figure 10a illustrates that the Monte Carlo sequence λ_{MC} is only slightly larger than $\lambda_{\text{G}^\star}$: the difference increases with the index i, and becomes substantial for ranges of i that are unlikely to be relevant in the scientific problem at hand.

Tables 1 and 2 in Service et al. (2014) describe a total of 14 variants as having an effect on HDL: two of these are for regions *FADS1* and *MVK* and the strength of the evidence in this specific dataset is quite weak (a marginal p-value of the order of 10^{-3}). Multiple effects are identified in regions *ABCA1, CEPT, LPL*, and *LIPL*. The results of the various "model selection" strategies we explored are in Figure 11, which reports the estimated values of the coefficients. The effect of the shrinkage induced by Lasso and SLOPE are evident: to properly compare effect sizes across methods it would be useful to resort to the two-step procedure that we used for the simulation described in Figure 2. Since our interest here is purely model selection, we report the coefficients directly as estimated by the ℓ_1 penalized procedures: this has the welcome side effect of increasing the spread of points in Figure 11, improving visibility.

Of the 14 variants described in Service et al. (2014), 8 are selected by all methods. The remaining 6 are all selected by at least some of the 8 methods we compared. There are an additional 5 variants that are selected by all methods but are not in the main list of findings in the original paper: four of these are rare variants, and one is an *array SNP* for a trait other than HDL. While none of these, therefore, was singularly analyzed for association in Service et al. (2014), they are in highlighted regions: one is in *MADD*, and the others in *ABCA1* and *CETP*, where the paper documents a plurality of signals.

Besides this core of common selections that correspond well to the original findings, there are notable differences among the 8 approaches we considered. The total number of selected variables ranges from 15—with BH on the p-values of the full model—to 119 with the cross-validated Lasso. It is not surprising that these methods would result in the extreme solutions: on the one hand, the p-values from the full model reflect the contribution of one variable given all the others, which are, however, not necessarily included in the models selected by other approaches; on the other hand, we have seen how the cross-validated Lasso tends to select a much larger number of variables and offers no control of FDR. In our case, the cross-validated Lasso estimates nonzero coefficients for 90 variables that are not selected by any other methods. Note that the number of variables selected by the cross-validated Lasso changes in different runs of the procedure, as implemented in `glmnet` with default parameters. It is quite reasonable to assume that a large number of these are false positives:

regions *G6PC2, PANK1, CRY2, MTNR1B*, where the Lasso-λ_{CV} selects some variants, have no documented association with lipid levels, and regions *CELSR2, GCKR, ABCG8*, and *NCAN* have been associated previously to total cholesterol and LDL, but not HDL. The other procedures that select some variants in any of these regions are the forward and backward greedy searches trying to optimize BIC, which have hits in *CELSR2* and *ABCG8*, and the BH on univariate p-value, which has one hit in *ABCG8*. SLOPE does not select any variant in regions not known to be associated with HDL. This is true also of the Lasso-λ_{Bonf} and BH on the p-values from the full model, but these miss respectively 2 and 6 of the variants described in the original paper, while SLOPE $\lambda_{\mathrm{G}^\star}$ misses only one of them.

Figure 12 focuses on the set of variants where there is some disagreement between the 8 procedures we considered, after eliminating the 90 variants selected only by the Lasso-λ_{CV}. In addition to recovering all except one of the variants identified in Service et al. (2014), and to the core of variants selected by all methods, SLOPE-$\lambda_{\mathrm{G}^\star}$ selects 3 rare variants and 3 common variants. While the rare variants were not singularly analyzed in the original study, they are in the two regions where aggregate tests highlighted the role of this type of variation. One is in *ABCA1* and the other two are in *CETP*, and they are both non-synonimous. Two of the three additional common variants are in *CETP* and one is in *MADD*: in addition to SLOPE, these are selected by Lasso–λ_{CV} and the marginal tests. One of the common variants and one rare variant in *CETP* are mentioned as a result of the limited foray in model selection in Service et al. (2014). SLOPE-λ_{MC} selects two less of these variants.

In order to get a handle on the effective FDR control of SLOPE in this setting, we resorted to simulations. We consider a number k of relevant variants ranging from 0 to 100, while concentrating on lower values. At each level, k columns of the design matrix were selected at random and assigned an effect of $\sqrt{2\log p}$ against a noise level σ set to 1. While analyzing the data with λ_{MC} and $\lambda_{\mathrm{G}^\star}$, we estimated σ from the full model in each run. Figure 10(b-c) reports the average FDP across 500 replicates and their standard error: the FDR of both λ_{MC} and $\lambda_{\mathrm{G}^\star}$ are close to the nominal levels for all $k \leq 100$.

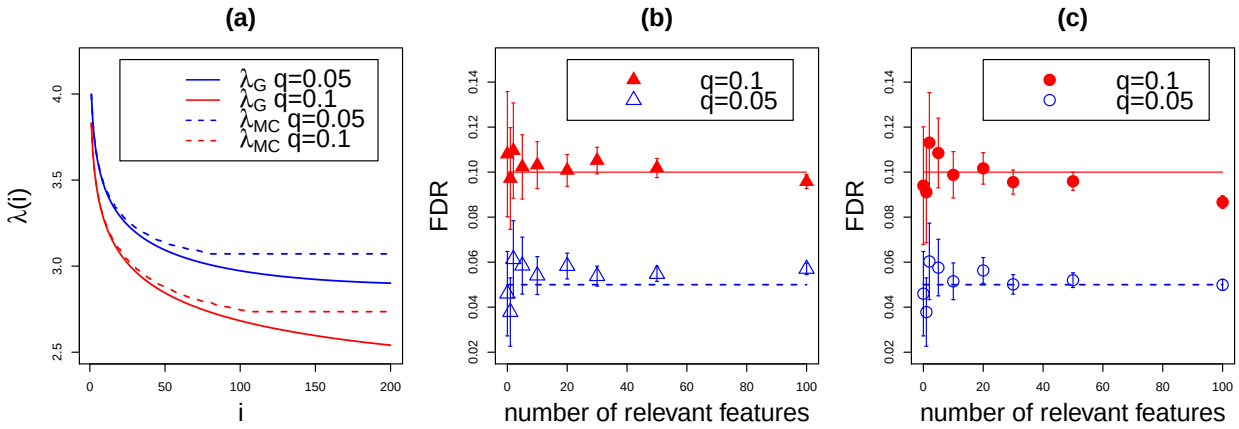


Figure 10: (a) Graphical representation of sequences λ_{MC} and λ_{G} for the variants design matrix. Mean FDP $\pm$ 2SE for SLOPE with (b) $\lambda_{\mathrm{G}^\star}$ and (c) λ_{MC} for the variants design matrix and $\beta_1 = \cdots = \beta_k = \sqrt{2\log p} \approx 3.65$, $\sigma = 1$.

In conclusion, the analysis with SLOPE confirms the results in Service et al. (2014), does not appear to introduce a large number of false positives, and hence it makes it easier to include in the final list of relevant variants a number of polymorphisms that are either directly highlighted in the original paper or in regions that were described as including a plurality of signals, but for which the original multi-step analysis did not allow to make a precise statement.

4 Discussion

The ease with which data are presently acquired has effectively created a new scientific paradigm: in addition to carefully designing experiments to test specific hypotheses, researchers often collect data first, leaving question formulation to a later stage. In this context, linear regression has increasingly been used to identify connections between one response and a large number p of possible explanatory variables. When $p \gg n$, approaches based on convex optimization have been particularly effective: an easily computable solution has the advantage of definitiveness and of reproducibility—another researcher, working on the same dataset, would obtain the same answer. Reproducibility of a scientific finding, or of the association between the outcome and the set of explanatory variables selected among many, however, is harder to achieve. Traditional tools such as p-values are often unhelpful in this context because of the difficulties of accounting for the effect of selection. In response, a great number of proposals (see e.g. Benjamini and Yekutieli (2005); Wasserman and Roeder (2009); Meinshausen et al. (2009); Berk et al. (2013); Zhang and Zhang (2013); van de Geer et al. (2014); Meinshausen and Bühlmann (2010); Efron (2011); Javanmard and Montanari (2013a,b); Bühlmann (2013); Lockhart et al. (2014))) present different approaches for controlling some measures of type I error in the context of variable selection. We here chose as a useful paradigm that of controlling the expected proportion of irrelevant variables among the selected ones. A similar goal of FDR control is pursued in Foygel-Barber and Candès (2014); G'Sell et al. (2013). While Foygel-Barber and Candès (2014) achieves exact FDR control in finite sample irrespective of the structure of the design matrix, this method, at least in the current implementation, is really best tailored for cases where $n > p$. The work in G'Sell et al. (2013) relies on p-values evaluated as in Lockhart et al. (2014), and is limited to the contexts where the assumptions in Lockhart et al. (2014) are met, including the assumption that all true regressors appear before the false regressors along the LASSO path. SLOPE controls FDR under orthogonal designs, and simulation studies also show that SLOPE can keep the FDR close to the nominal level when $p > n$ and the true model is sparse, while offering large power and accurate prediction. This is, of course, only a starting point and many open problems remain.

First, while our heuristics for the choice of the λ sequence allows to keep FDR under control for Gaussian designs and other random design matrices (more examples are provided in Bogdan et al. (2013)), it is by no means a definite solution. Further theoretical research is needed to identify the sequences λ, which would provably control FDR for these designs and other typical design matrices.

Second, just as in the BH procedure where the test statistics are compared with fixed critical values, we have only considered in this paper fixed values of the regularizing sequence $\{\lambda_i\}$. It would be interesting to know whether it is possible to select such parameters in a data-driven fashion as to achieve desirable statistical properties. For the simpler Lasso problem for instance, an important question is whether it is possible to select λ on the Lasso path as to control the FDR. In the case where $n \geq p$ a method to obtain this goal was recently proposed in Foygel-Barber and Candès (2014). It would be of great interest to know if similar positive theoretical results can be

obtained for SLOPE, in perhaps restricted sparse settings.

Third, our research points the limits of signal sparsity which can be handled by SLOPE. Such limitations are inherent to ℓ_1 convex optimization methods and also pertain to Lasso. Some discussion on the minimal FDR which can be obtained with Lasso under Gaussian designs is provided in Bogdan et al. (2013), while new evocative results on adaptive versions of Lasso are on the way.

Fourth, we illustrated the potential of SLOPE for multiple testing with positively correlated test statistics. In our simple ANOVA model, SLOPE controls FDR even when the unknown variance components are replaced with their estimates. It remains an open problem to theoretically describe a possibly larger class of unknown covariance matrices for which SLOPE can be used effectively.

In conclusion, we hope that the work presented so far would convince the reader that SLOPE is an interesting convex program with promising applications in statistics and motivates further research.

Acknowledgements

We would like to thank the Editor, Professor Karen Kafadar, the Associate Editor and two reviewers for many constructive suggestions, which led to a substantial improvement of this article.

E. C. is partially supported by AFOSR under grant FA9550-09-1-0643, by ONR under grant N00014-09-1-0258 and by a gift from the Broadcom Foundation. M. B. was supported by the Fulbright Scholarship, NSF grant NSF 1043204 and the European Unions 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552. E. v.d.B. was supported by National Science Foundation Grant DMS 0906812 (American Reinvestment and Recovery Act). C. S. is supported by NIH grants HG006695 and MH101782. W. S. is supported by a General Wang Yaowu Stanford Graduate Fellowship. We thank the authors of Service et al. (2014) for letting us use their data during the completion of dbGaP release. E. C. would like to thank Stephen Becker for all his help in integrating the sorted ℓ_1 norm software into TFOCS. M. B. would like to thank David Donoho and David Siegmund for encouragement and Hatef Monajemi for helpful discussions. We are very grateful to Lucas Janson for suggesting the acronym SLOPE, and to Rina Foygel Barber and Julie Josse for useful comments about an early version of the manuscript.

[id-suppA] Supplement to "SLOPE – Adaptive Variable Selection via Convex Optimization" [doi]COMPLETED BY THE TYPESETTER We provide proofs of some technical results discussed in the text.

References

F. Abramovich and Y. Benjamini. Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *In Wavelets and Statistics, Lecture Notes in Statistics 103, Antoniadis*, pages 5–14. Springer-Verlag, 1995.

F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19: 716–723, 1974. ISSN 0018-9286. System identification and time-series analysis.

R. E. Barlow, D. J. Bartholomew, J-M. Bremner, and H.D. Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression.* Wiley New York, 1972.

P. Bauer, B. M. Pötscher, and P. Hackl. Model selection by multiple test procedures. *Statistics*, 19:39–44, 1988.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2:183–202, March 2009.

S. Becker, E. J. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, August 2011.

Y. Benjamini and Y. Gavrilov. A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.*, 3(1):179–198, 2009. ISSN 1932-6157. doi: 10.1214/08-AOAS194. URL `http://dx.doi.org/10.1214/08-AOAS194`.

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)*, 57(1):289–300, 1995.

Y. Benjamini and D. Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, 100(469):71–93, 2005. ISSN 0162-1459. doi: 10.1198/016214504000001907. URL `http://dx.doi.org/10.1198/016214504000001907`. With comments and a rejoinder by the authors.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013. ISSN 0090-5364. doi: 10.1214/12-AOS1077. URL `http://dx.doi.org/10.1214/12-AOS1077`.

M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1-3):425–439, 1990.

L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. ISSN 1435-9855.

M. Bogdan, J. K. Ghosh, and M Żak-Szatkowska. Selecting explanatory variables with the modified version of Bayesian information criterion. *Quality and Reliability Engineering International*, 24:627–641, 2008.

M. Bogdan, A. Chakrabarti, F. Frommlet, and J. K. Ghosh. Asymptotic Bayes optimality under sparsity of some multiple testing procedures. *Annals of Statistics*, 39:1551–1579, 2011.

M. Bogdan, E. van den Berg, W. Su, and E. J. Candès. Statistical estimation and testing via the ordered ℓ_1 norm. arXiv:1310.1969v2, 2013.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. Supplement to "slope – adaptive variable selection via convex optimization". 2015.

H. D. Bondell, X., and B. J. Reich. Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics*, pages 115–123, 2008.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013. ISSN 1350-7265. doi: 10.3150/12-BEJSP11. URL `http://dx.doi.org/10.3150/12-BEJSP11`.

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

E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *J. Fourier Anal. Appl.*, 14:877–905, 2008.

J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

B. Efron. Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.*, 106(496):1602–1614, 2011. ISSN 0162-1459. doi: 10.1198/jasa.2011.tm11181. URL `http://dx.doi.org/10.1198/jasa.2011.tm11181`.

D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4): 1947–1975, 1994. ISSN 0090-5364.

D. P. Foster and R. A. Stine. Local asymptotic coding and the minimum description length. *IEEE Transactions on Information Theory*, 45(4):1289–1293, 1999.

R. Foygel-Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. arXiv:1404.5609, to appear in *Ann. Statist.*, 2014.

F. Frommlet and M. Bogdan. Some optimality properties of FDR controlling rules under sparsity. *Electronic Journal of Statistics*, 7:1328–1368, 2013.

F. Frommlet, F. Ruhaltinger, P. Twaróg, and M. Bogdan. A model selection approach to genome wide association studies. *Computational Statistics & Data Analysis*, 56:1038–1051, 2012.

S.J. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied Mathematics and Optimization*, 12(1):247–270, 1984. ISSN 0095-4616.

M. Grazier G'Sell, T. Hastie, and R. Tibshirani. False variable selection rates in regression. *arXiv:1302.2303*, 2013.

Y. Ingster. Minimax detection of a signal for l^n-balls. *Math. Methods Statist.*, 7:401–428, 1999.

A. Javanmard and A. Montanari. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *ArXiv e-prints*, June 2013a.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *ArXiv e-prints*, 2013b.

J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964.

R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of Statistics*, 42:413–468, 2014.

C. L. Mallows. Some comments on c_p. *Technometrics*, 15(2):661–676, 1973.

N. Meinshausen. Relaxed lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007.

N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473, 2010. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2010.00740.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x`.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. `http://www.ecore.be/DPs/dp_1191313936.pdf`, 2007. CORE discussion paper.

N. Parikh and S. Boyd. Proximal algorithms. In *Foundations and Trends in Optimization*, volume 1, pages 123–231. 2013.

S. K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, pages 239–257, 2002.

S. K. Service, T. M. Teslovich, C. Fuchsberger, V. Ramensky, P. Yajnik, D. C. Koboldt, D. E. Larson, Q. Zhang, L. Lin, R. Welch, L. Ding, M. D. McLellan, M. O'Laughlin, C. Fronick, L. L. Fulton, V. Magrini, A. Swift, P. Elliott, M. R. Jarvelin, M. Kaakinen, M. I. McCarthy, L. Peltonen, A. Pouta, L. L. Bonnycastle, F. S. Collins, N. Narisu, H. M. Stringham, J. Tuomilehto, S. Ripatti, R. S. Fulton, C. Sabatti, R. K. Wilson, M. Boehnke, and N. B. Freimer. Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet.*, 10(1):e1004147, Jan 2014.

N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models (with discussion). *Test*, 19:209–285, 2010.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

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

R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *J. Roy. Statist. Soc. B*, 55:757–796, 1999.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014.

L. Wasserman and K. Roeder. High-dimensional variable selection. *Annals of Statistics*, pages 2178–2201, 2009.

Z. Wu and H. H. Zhou. Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1-2):165–191, 2013.

X. Zeng and M. Figueiredo. Decreasing weighted sorted l1 regularization. *IEEE Signal Processing Letters*, pages 1240–1244, 2014.

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*, pages 217–242, 2013.

L. Zhong and J. Kwok. Efficient sparse modeling with automatic feature grouping. *IEEE Trans. Neural Netw. Learn. Syst.*, pages 1436–1447, 2012.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101: 1418–1429, 2006.

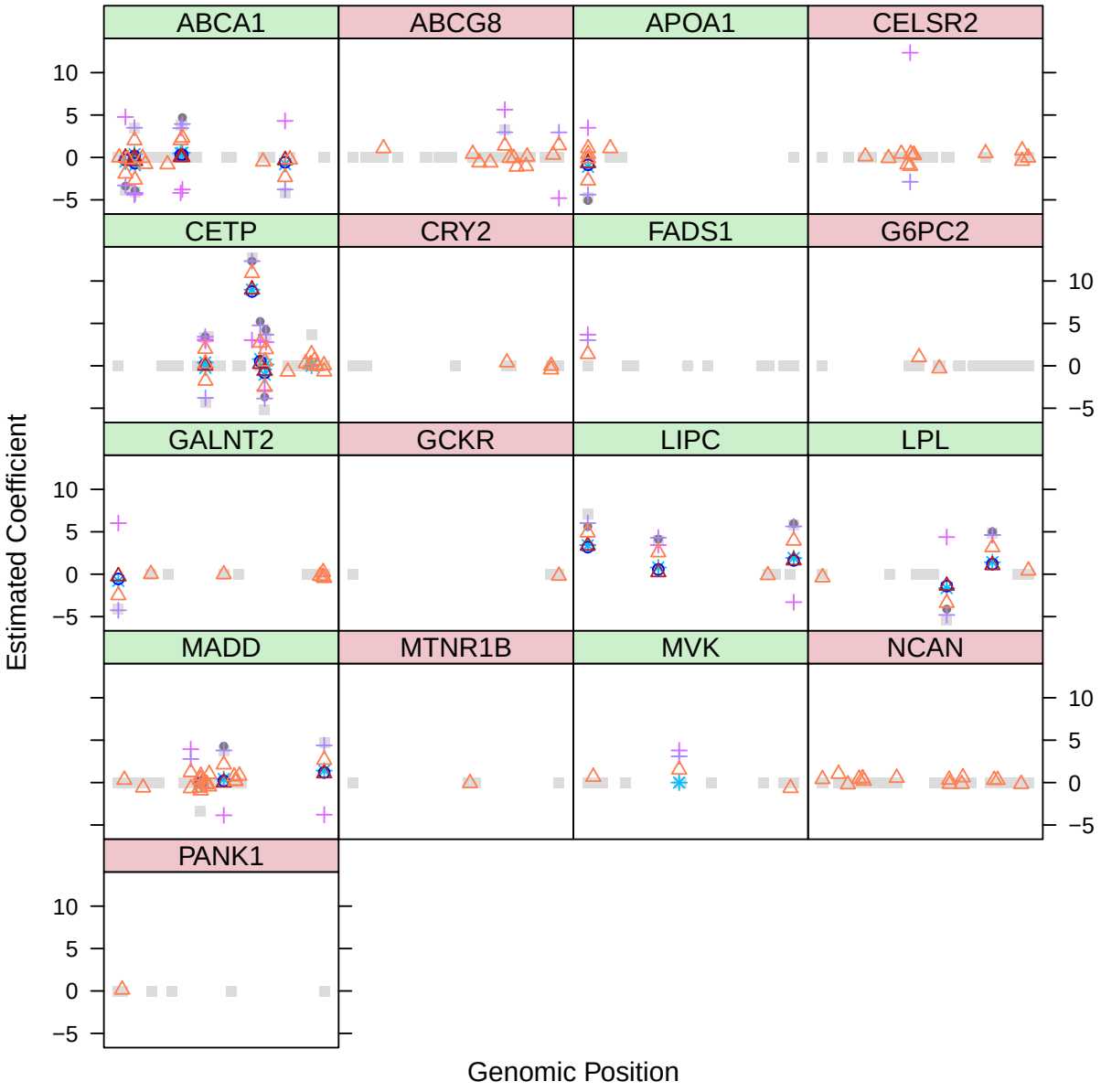


Figure 11: Estimated effects on HDL for variants in 17 regions. Each panel corresponds to a region and is identified by the name of a gene in the region, following the convention in Service et al. (2014). Regions with (without) previously reported association to HDL are on green (red) background. On the x-axis variants position in base-pairs along their respective chromosomes. On the y-axis estimated effect according to different methodologies. With the exception of marginal tests—which we use to convey information on the number of variables and indicated with light gray squares—we report only the value of non zero coefficients. The rest of the plotting symbols and color convention is as follows: dark gray bullet—BH on p-values from full model; magenta cross—forward BIC; purple cross—backwards BIC; red triangle—Lasso–λ_{Bonf}; orange triangle—Lasso–λ_{CV}; cyan star—SLOPE–$\lambda_{\text{G}^\star}$; black circle—SLOPE with λ defined with Monte Carlo strategy.

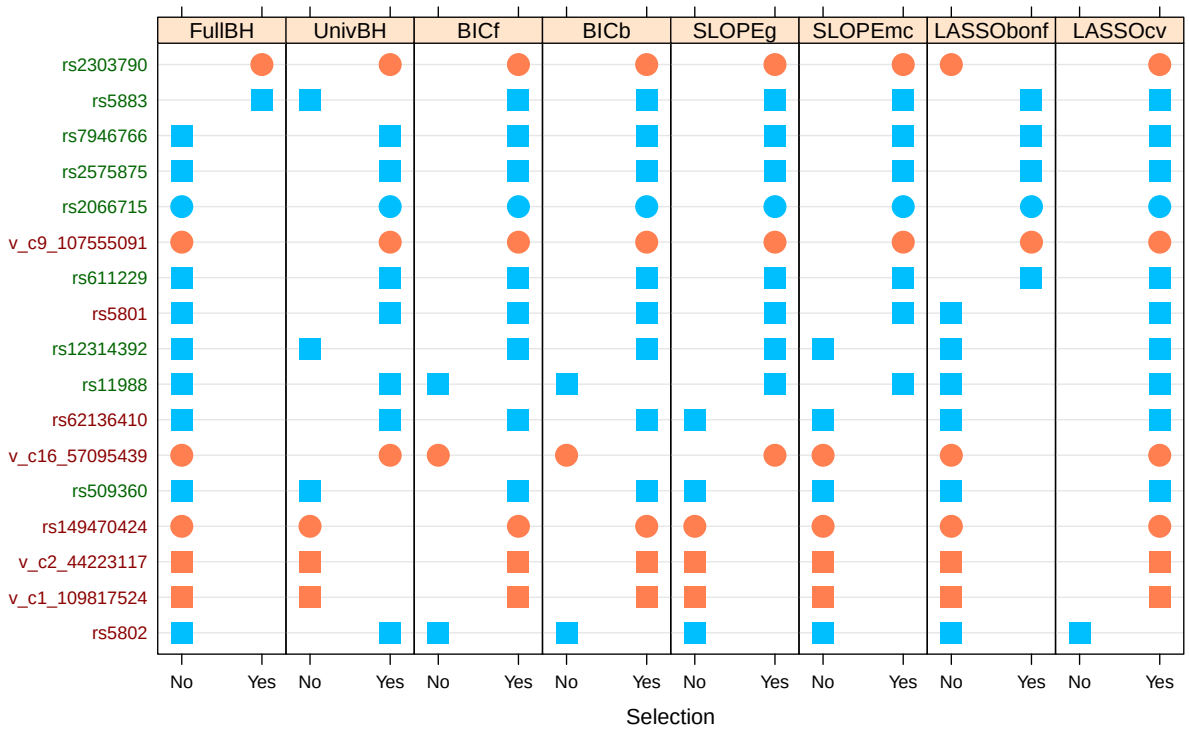


Figure 12: Each row corresponds to a variant in the set differently selected by the the compared procedures, indicated by columns. Orange is used to represent rare variants and blue common ones. Squares indicate synonymous (or non coding variants) and circle non-synonimous ones. Variants are ordered according to the frequency with which they are selected. Variants with names in green are mentioned in Service et al. (2014) as to have an effect on LDL, while variants with names in red are not (if a variant was not in dbSNP build 137, we named it by indicating chromosome and position, following the convention in Service et al. (2014)).