Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Group SLOPE - adaptive selection of groups of predictors

Damian Brzyski, Alexej Gossmann, Weijie Su & Małgorzata Bogdan

To cite this article: Damian Brzyski, Alexej Gossmann, Weijie Su & Małgorzata Bogdan (2018): Group SLOPE - adaptive selection of groups of predictors, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1411269

To link to this article: https://doi.org/10.1080/01621459.2017.1411269

Group SLOPE - adaptive selection of groups of predictors [1]

Damian Brzyski[a,b], Alexej Gossmann[c], Weijie Su[d], Małgorzata Bogdan[e]

[a] *Department of Epidemiology and Biostatistics, Indiana University, Bloomington, IN 47405, USA*

[b] *Institute of Mathematics, Jagiellonian University, 30-348 Cracow, Poland*

[c] *Bioinnovation PhD Program, Tulane University, New Orleans, LA 70118, USA*

[d] *Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104, USA*

[e] *Institute of Mathematics, University of Wroclaw, 50-384 Wroclaw, Poland*

Key words: Asymptotic Minimax, False Discovery Rate, Group selection, Model Selection, Multiple Regression , SLOPE

Abstract

Sorted L-One Penalized Estimation (SLOPE, Bogdan et al., 2013, 2015) is a relatively new convex optimization procedure which allows for adaptive selection of regressors under sparse high dimensional designs. Here we extend the idea of SLOPE to deal with the situation when one aims at selecting whole groups of explanatory variables instead of single regressors. Such groups can be formed by clustering strongly correlated predictors or groups of dummy variables corresponding to different levels of the same qualitative predictor. We formulate the respective convex optimization problem, gSLOPE (group SLOPE), and propose an efficient algorithm for its solution. We also define a notion of the group false discovery rate (gFDR) and provide a choice of the sequence of tuning parameters for gSLOPE so that gFDR is provably controlled at a prespecified level if the groups of variables are orthogonal to each other. Moreover, we prove that the resulting procedure adapts to unknown sparsity and is asymptotically minimax

[1] An earlier version of the paper appeared on arXiv.org in November 2015: arXiv:1511.09078

with respect to the estimation of the proportions of variance of the response variable explained by regressors from different groups. We also provide a method for the choice of the regularizing sequence when variables in different groups are not orthogonal but statistically independent and illustrate its good properties with computer simulations. Finally, we illustrate the advantages of gSLOPE in the context of Genome Wide Association Studies. `R` package `grpSLOPE` with an implementation of our method is available on CRAN.

1 Introduction

Consider the classical multiple regression model of the form

$$y = X\beta + z, \tag{1.1}$$

where y is the n dimensional vector of values of the response variable, X is the n by p experiment (design) matrix and $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. We assume that y and X are known, while β is unknown. In many applications the purpose of the statistical analysis is to recover the support of β, which identifies the set of important regressors. Here, the true support corresponds to truly relevant variables (i.e. variables which have impact on observations). Common procedures to solve this model selection problem rely on minimization of some objective function consisting of the weighted sum of two components: first term responsible for the goodness of fit and second term penalizing the model complexity. Among such procedures one can mention classical model selection criteria like the Akaike Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), where the penalty depends on the number of variables included in the model, or LASSO (Tibshirani, 1996), where the penalty depends on the ℓ_1 norm of regression coefficients. The main advantage of LASSO over classical model selection criteria is that it is a convex optimization problem and, as

such, it can be easily solved even for very large design matrices.

LASSO solution is obtained by solving the optimization problem

$$\underset{b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} \left\| y - Xb \right\|^2 + \lambda_L \|b\|_1 \right\}, \tag{1.2}$$

where λ_L is a tuning parameter defining the trade-off between the model fit and the sparsity of solution. In practical applications the selection of good λ_L might be very challenging. For example it has been reported that in high dimensional settings the popular cross-validation typically leads to detection of a large number of false regressors (see e.g. Bogdan et al., 2015). The general rule is that when one reduces λ_L, then LASSO can identify more elements from the true support (true discoveries) but at the same time it generates more false discoveries. In general the numbers of true and false discoveries for a given λ_L depend on unknown properties on the data generating mechanism, like the number of true regressors and the magnitude of their effects. A very similar problem occurs when selecting thresholds for individual tests in the context of multiple testing. Here it was found that the popular Benjamini-Hochberg rule (BH Benjamini and Hochberg, 1995), aimed at control of the False Discovery Rate (FDR), adapts to the unknown data generating mechanism and has some desirable optimality properties under a variety of statistical settings (see e.g. Abramovich et al., 2006; Bogdan et al., 2011; Neuvial and Roquain, 2012; Frommlet and Bogdan, 2013). The main property of this rule is that it relaxes the thresholds along the sequence of test statistics, sorted in the decreased order of magnitude. Recently the same idea was used in a new generalization of LASSO, named SLOPE (Sorted L-One Penalized Estimation, Bogdan et al., 2013, 2015). Instead of the ℓ_1 norm (as in LASSO case), the method uses FDR control properties of J_λ norm, defined as follows; for sequence $\{\lambda\}_{i=1}^p$ satisfying $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ and $b \in \mathbb{R}^p$, $J_\lambda(b) := \sum_{i=1}^p \lambda_i |b|_{(i)}$, where $|b|_{(1)} \geq \ldots \geq |b|_{(p)}$ is the vector of sorted absolute values of coordinates of b. SLOPE is the solution to a convex

optimization problem

$$\underset{b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} \left\| y - Xb \right\|^2 + J_\lambda(b) \right\}, \tag{1.3}$$

which clearly reduces to LASSO for $\lambda_1 = \ldots = \lambda_p =: \lambda_L$. Similarly as in classical model selection, the support of the solution defines the subset of variables estimated as relevant. In (Bogdan et al., 2013, 2015) it is shown that when the sequence λ corresponds to the decreasing sequence of thresholds for BH then SLOPE controls FDR under orthogonal designs, i.e. when $X^T X = \mathbf{I}_n$. Moreover, in (Su and Candès, 2016) it is proved that SLOPE with this sequence of tuning parameters adapts to unknown sparsity and is asymptotically minimax under orthogonal and random Gaussian designs.

In the sequence of examples presented in (Bogdan et al., 2013, 2015; Brzyski et al., 2017) it was shown that SLOPE has very desirable properties in terms of FDR control in the case when regressor variables are weakly correlated. While there exist other interesting approaches which allow to control FDR under correlated designs (e.g. Barber and Candès, 2015), the efforts to prevent detection of false regressors which are strongly correlated with true ones inevitably lead to a loss of power. An alternative approach to deal with strongly correlated predictors is to simply give up the idea of distinguishing between them and include all of them into the selected model as a group. This leads to the problem of group selection in linear regression, extensively investigated and applied in many fields of science. In many of these applications the groups are selected not only due to the strong correlations but also by taking into account the problem-specific scientific knowledge. It is also common to cluster dummy variables corresponding to different levels of qualitative predictors.

Probably the most well known convex optimization method for selection of groups of explanatory variables is the group (gLASSO Bakin, 1999). For a fixed tuning parameter, $\lambda_{gL} > 0$, the gLASSO estimate is most frequently (e.g. Yuan and Lin, 2006; Simon et al., 2013) defined as a solution to optimization problem

$$\underset{b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^{m} X_{I_i} b_{I_i} \right\|_2^2 + \sigma \lambda_{gL} \sum_{i=1}^{m} \sqrt{|I_i|} \|b_{I_i}\|_2 \right\}, \tag{1.4}$$

where the sets $I_1, \ldots, I_m$ form a partition of the set $\{1, \ldots, p\}$, $|I_i|$ denotes the number of elements in set I_i, X_{I_i} is the submatrix of X composed of columns indexed by I_i and b_{I_i} is the restriction of b to indices from I_i. The method introduced in this article is, however, closer to the alternative version of gLASSO, in which penalties are imposed on $\|X_{I_i} b_{I_i}\|_2$ rather than $\|b_{I_i}\|_2$. This method was formulated in (Simon and Tibshirani, 2013), where the authors defined an estimate of β by

$$\beta^{\mathrm{gL}} := \underset{b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^{m} X_{I_i} b_{I_i} \right\|_2^2 + \sigma \lambda_{gL} \sum_{i=1}^{m} \sqrt{|I_i|} \|X_{I_i} b_{I_i}\|_2 \right\}, \tag{1.5}$$

with the condition $\|X_{I_i} \beta^{\mathrm{gL}}_{I_i}\|_2 > 0$ serving as a group relevance indicator.

Similarly as in the context of regular model selection, the properties of gLASSO strongly depend on the shrinkage parameter λ_{gL}, whose optimal value is the function of unknown parameters of true data generating mechanism. Thus, a natural question arises of whether the idea of SLOPE can be used for construction of a similar adaptive procedure for the group selection. To answer this query in this paper we define and investigate the properties of the group SLOPE (gSLOPE). We formulate the respective optimization problem and provide the algorithm for its solution. We also define the notion of the group FDR (gFDR), and provide the theoretical choice of the sequence of regularization parameters, which guarantees that gSLOPE controls gFDR in the situation when variables in different groups are orthogonal to each other. Moreover, we prove that the resulting procedure adapts to unknown sparsity and is asymptotically minimax with respect to the estimation of the proportions of variance of the response variable explained by regressors from different groups. Additionally, we provide a way of constructing the sequence of regularization parameters under the assumption that the regressors from distinct groups are independent and use computer simulations to show that it allows to control gFDR. Good properties of group SLOPE are illustrated

using the practical example of Genome Wide Association Study. `R` package `grpSLOPE` with an implementation of our method is available on CRAN. All scripts used in simulations as well as in real data analysis are available at `https://github.com/dbrzyski/gSLOPE`. This repository contains also R scripts which were used to generate article figures.

2 Group SLOPE

2.1 Formulation of the optimization problem

Let the design matrix X belong to the space $M(n,p)$ of matrices with n rows and p columns. Furthermore, suppose that $I = \{I_1, \ldots, I_m\}$ is some partition of the set $\{1, \ldots, p\}$, i.e. I_i's are nonempty sets, $I_i \cap I_j = \emptyset$ for $i \neq j$ and $\bigcup I_i = \{1, \ldots, p\}$. We will consider the linear regression model with m groups of the form

$$y = \sum_{i=1}^{m} X_{I_i} \beta_{I_i} + z, \tag{2.1}$$

where X_{I_i} is the submatrix of X composed of columns indexed by I_i and β_{I_i} is the restriction of β to indices from the set I_i. We will use notation $l_1, \ldots, l_m$ to refer to the ranks of submatrices $X_{I_1}, \ldots, X_{I_m}$. To simplify notation later, we will assume that $l_i > 0$ (i.e. there is at least one nonzero entry of X_{I_i} for all i). Besides this, X may be an arbitrary matrix, in particular linear dependence inside each of the submatrices X_{I_i} is allowed.

In this article we will treat the value $\|X_{I_i}\beta_{I_i}\|_2$ as a measure of an impact of i^{th} group on the response and we will say that the group i is truly relevant if and only if $\|X_{I_i}\beta_{I_i}\|_2 > 0$. Thus our task of the identification of the relevant groups is equivalent with finding the support of the vector $[\![\beta]\!]_{I,X} := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2\right)^{\mathsf{T}}$.

To estimate the nonzero coefficients of $[\![\beta]\!]_{I,X}$, we will use a new penalized method,

namely group SLOPE (gSLOPE). For a given nonincreasing sequence of nonnegative tuning parameters, $\lambda_1, \ldots, \lambda_m$, a given sequence of positive weights, $w_1, \ldots, w_m$, and a design matrix, X, the gSLOPE estimator of regression coefficients, β^{gS}, is defined as any solution to the optimization problem

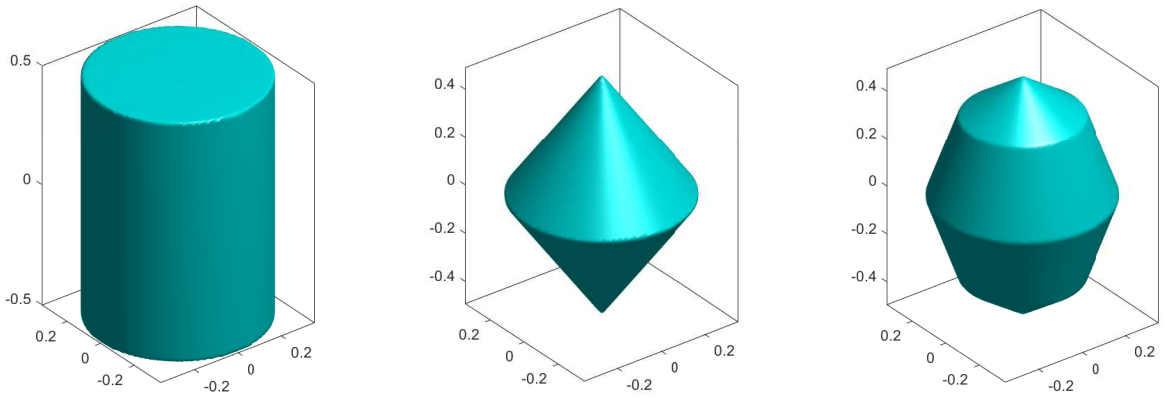
$$\beta^{\mathrm{gS}} := \underset{b}{\operatorname{argmin}} \quad \Big\{ \frac{1}{2} \big\| y - Xb \big\|_2^2 + \sigma J_\lambda\big(W [\![b]\!]_{I,X} \big) \Big\}, \tag{2.2}$$

where W is a diagonal matrix with $W_{i,i} := w_i$, for $i = 1, \ldots, m$. The estimate of $[\![\beta]\!]_{I,X}$ support is simply defined by the indices corresponding to nonzeros of $[\![\beta^{\mathrm{gS}}]\!]_{I,X}$.

It is easy to see that when one considers p groups containing only one variable (i.e. singleton groups situation), then taking all weights equal to one reduces (2.2) to SLOPE (1.3). On the other hand, taking $w_i = \sqrt{|I_i|}$ and putting $\lambda_1 = \ldots = \lambda_m =: \lambda_{gL}$, immediately gives gLASSO problem (1.5) with the smoothing parameter λ_{gL}. The gSLOPE could be therefore treated both: as the extension to SLOPE, and the extension to group LASSO.

As shown in Appendix B the function $J_{\lambda,I,W,X}(b) := J_\lambda\big(W [\![b]\!]_{I,X} \big)$ is a seminorm and becomes a norm when the design matrix X is of the full rank. Figure 1 illustrates how the shape of the unit ball in the norm $J_{\lambda,I,W}(b) := J_\lambda\big(W [\![b]\!]_I \big)$ depends on the selection of the λ sequence. In this example $p = 3$, $m = 2$, $I_1 := \{1, 2\}$ and $I_2 := \{3\}$. In case when only the first coefficient in the λ sequence is larger than zero $J_{\lambda,I,W}(b) = \lambda_1 \max_{i \in \{1,2\}} w_i ||\beta_{I_i}||_2$, and the corresponding ball takes form of the cylinder. The privileged solutions occur on the "edges" of this cylinder and have the same weighted group effects for both groups (i.e. $w_1 \sqrt{\beta_1^2 + \beta_2^2} = w_2 |\beta_3|$). When $\lambda_1 = \lambda_2 > 0$ then the the ball takes the form of the "spinning top", with "edges" occuring when at least one group effect is equal to zero. Then the group SLOPE reduces to the group LASSO and has a tendency to select a sparse solution. When $\lambda_1 > \lambda_2 > 0$ the corresponding ball has both types of edges and encourages the dimensionality reduction in both ways: by

inducing the sparsity and making some of the weighted group effects to be equal to each other.



(a) $\lambda_1 = 2,\ \lambda_2 = 0$ (b) $\lambda_1 = 2,\ \lambda_2 = 2$ (c) $\lambda_1 = 2,\ \lambda_2 = 1$

Figure 1: Unit balls of $J_{\lambda,I,W}$ norm for different λ. Weights w_1 and w_2 are equal to $\sqrt{2}$ and 1. The edges in (a) correspond to the same weighted group effects, i.e. $w_1\sqrt{\beta_1^2+\beta_2^2} = w_2|\beta_3| = 0.5$; all edges in (b) contain at least one zero group effect (gLASSO); in (c) both types of edges appear.

Now, let us again consider an arbitrary $m > 0$, define $\widetilde{p} = l_1 + \ldots + l_m$ and examine the following partition, $\mathbb{I} = \{\mathbb{I}_1, \ldots, \mathbb{I}_m\}$, of the set $\{1, \ldots, \widetilde{p}\}$

$$\mathbb{I}_1 := \big\{1, \ldots, l_1\big\}, \quad \mathbb{I}_2 := \big\{l_1 + 1, \ldots, l_1 + l_2\big\}, \quad \ldots, \quad \mathbb{I}_m := \Big\{\sum_{j=1}^{m-1} l_i + 1, \ldots, \sum_{j=1}^{m} l_i\Big\}.$$

Observe that each X_{I_i} can be represented as $X_{I_i} = U_i R_i$, where U_i is a matrix with l_i orthogonal columns of a unit l_2 norm, whose span coincides with the space spanned by the columns of X_{I_i}, and R_i is the corresponding matrix of a full row rank. Define n by l matrix $\widetilde{X}$ by putting $\widetilde{X}_{\mathbb{I}_i} := U_i$ for $i = 1, \ldots, m$. Now observe that after defining vector ω by conditions $\omega_{\mathbb{I}_i} := R_i b_{I_i}$ for $i \in \{1, \ldots, m\}$, we immediately obtain

$$\begin{gathered} Xb = \textstyle\sum_{i=1}^m X_{I_i} b_{I_i} = \sum_{i=1}^m U_i R_i b_{I_i} = \sum_{i=1}^m \widetilde{X}_{\mathbb{I}_i} \omega_{\mathbb{I}_i} = \widetilde{X}\omega, \\ \Big(\big[\!\big[b\big]\!\big]_{I,X}\Big)_i = \|X_{I_i} b_{I_i}\|_2 = \|R_i b_{I_i}\|_2 = \|\omega_{\mathbb{I}_i}\|_2 \end{gathered} \tag{2.3}$$

and for $\big[\!\big[\omega\big]\!\big]_{\mathbb{I}} := \big(\|\omega_{\mathbb{I}_1}\|_2, \ldots, \|\omega_{\mathbb{I}_m}\|_2\big)^{\mathsf{T}}$ the problem (2.2) can be equivalently presented

in the form

$$\omega^{\mathrm{gS}} := \underset{\omega}{\operatorname{argmin}} \quad \left\{ \frac{1}{2} \left\| y - \widetilde{X}\omega \right\|_2^2 + \sigma J_\lambda\big(W[\![\omega]\!]_{\mathbb{I}}\big) \right\}, \tag{2.4}$$

where ω^{gS} and β^{gS} are linked via conditions $\omega^{\mathrm{gS}}_{\mathbb{I}_i} := R_i \beta^{\mathrm{gS}}_{I_i},\ i = 1, \ldots, m$. Therefore to identify the relevant groups and estimate their group effects it is enough to solve the optimization problem (2.4). We will say that (2.4) is the standardized version of the problem (2.2).

Remark 2.1. *The formulation of the group SLOPE was proposed independently in (Brzyski et al., 2015) (earlier version of this article) and in (Gossmann et al., 2015). In (Gossmann et al., 2015) only the case when the weights w_i are equal to the square root of the group size is considered and penalties are imposed directly on $\|\beta_{I_i}\|_2$ rather than on group effects $\|X_{I_i}\beta_{I_i}\|_2$. This makes the method of (Gossmann et al., 2015) dependent on scaling or rotations of variables in a given group. In comparison to (Gossmann et al., 2015), where a Monte Carlo approach for estimating the regularizing sequence was proposed, our article lays theoretical foundations and provides the guidelines for the choice of the sequence of smoothing parameters, so gSLOPE can control FDR and has desired estimation properties.*

2.2 Numerical algorithm

We will at first show that the problem of solving (2.4) can be easily reduced to the situation when W is the identity matrix. For this aim we define a diagonal matrix M such that for $j \in \mathbb{I}_i$ $M_{j,j} := w_i^{-1}$. Then observe that

$$J_{\lambda,\mathbb{I},W}(\omega) = J_\lambda\big(W[\![\omega]\!]_{\mathbb{I}}\big) = J_\lambda\big([\![M^{-1}\omega]\!]_{\mathbb{I}}\big) = J_{\lambda,\mathbb{I},\mathbf{I}_{\tilde{p}}}\big(M^{-1}\omega\big) =: J_{\lambda,\mathbb{I}}\big(M^{-1}\omega\big). \tag{2.5}$$

Since M is nonsingular, we can substitute $\eta := M^{-1}\omega$ and consider equivalent formulation of (2.4), $\eta^* := \operatorname{argmin}_\eta \left\{ \frac{1}{2}\|y - \widetilde{X}M\eta\|_2^2 + J_{\sigma\lambda,\mathbb{I}}\big(\eta\big) \right\}$, and recover ω^{gS} as $\omega^{\mathrm{gS}} = M\eta^*$. This allows to recast gSLOPE as a problem with unit weights.

Now, the above problem is of the form $\underset{b}{\text{minimize}}\ \{\underbrace{\|y-\mathcal{X}b\|_2^2/2}_{g(b)} + \underbrace{J_{\lambda,\mathbb{I}}(b)}_{h(b)}\}$, where g and h are convex functions and g is differentiable. There exist efficient methods, namely *proximal gradient algorithms*, which could be applied to find numerical solution in such situation. To design efficient algorithms, however, h must be prox-capable, meaning that there is known fast algorithm for computing the proximal operator for h,

$$\text{prox}_{th}(u) := \underset{b}{\text{argmin}} \left\{ \frac{1}{2t}\|u-b\|_2^2 + h(b) \right\}, \tag{2.6}$$

for each $u \in \mathbb{R}^{\widetilde{p}}$ and $t > 0$.

To derive the proximal operator for the group SLOPE, we at first assume without the loss of generality that $\sigma = 1$. Now, we need to find the algorithm to minimize $\frac{1}{2t}\|u-b\|_2^2 + J_{\lambda,\mathbb{I}}(b)$, for any $u \in \mathbb{R}^{\widetilde{p}}$ and $t > 0$, which is equivalent to finding the numerical solution to the problem

$$\text{prox}_J(u) := \underset{b}{\text{argmin}} \left\{ \frac{1}{2}\|u-b\|_2^2 + J_{\tilde{\lambda},\mathbb{I}}(b) \right\}, \quad \text{for} \quad \tilde{\lambda} := t\lambda. \tag{2.7}$$

As discussed in Appendix D, this problem can be solved in two steps

$$\begin{cases} c^* := \text{argmin}_{c \in R^m} \left\{ \frac{1}{2}\left\| [\![u]\!]_{\mathbb{I}} - c \right\|_2^2 + J_{\tilde{\lambda}}(c) \right\} \\ \left(\text{prox}_J(u)\right)_{\mathbb{I}_i} = c_i^* \left(\|u_{\mathbb{I}_i}\|_2\right)^{-1} u_{\mathbb{I}_i}, \quad i = 1, \dots, m \end{cases} \tag{2.8}$$

Consequently, calculating $\text{prox}_J(u)$ in fact reduces to identifying c^*, which can be efficiently done using the fast prox algorithm for regular SLOPE, provided e.g. in (Bogdan et al., 2013, 2015).

After defining the proximal operator, the solution to the gSLOPE can be obtained by the Procedure 1. There exist many ways in which t_i's can be selected to ensure that $f(b^{(k)})$ converges to the optimal value (see e.g. Beck and Teboulle, 2009; Tseng, 2008). In our `R` package `grpSLOPE` available on CRAN (The Comprehensive R Archive Network) the accelerated proximal gradient method known as FISTA (Beck and Teboulle, 2009) is applied, which uses the specific procedure for choosing steps sizes, to achieve

Procedure 1 Proximal gradient algorithm

input: $b^{[0]} \in \mathbb{R}^{\widetilde{p}}$, k=0
while (Stopping criteria are not satisfied) **do**
1. $b^{[k+1]} = \text{prox}_{t_k h}\left(b^{[k]} - t_k \mathcal{X}^{\mathsf{T}}(\mathcal{X} b^{[k]} - y)\right)$;
2. $k \leftarrow k + 1$.

end while

a fast convergence rate. To derive the proper stopping criterion, we have considered the dual problem to gSLOPE and employed the strong duality property. The detailed description of the dual norm, conjugate of grouped sorted l_1 norm and the stopping criterion are provided in the Appendix C.

2.3 Group FDR

Group SLOPE is designed to select groups of variables, which might be very strongly correlated within a group or even linearly dependent. In this context we do not intend to identify single important predictors but rather want to point at the groups which contain at least one true regressor. To theoretically investigate the properties of gSLOPE in this context we now introduce the respective notion of group FDR (gFDR).

Definition 2.2. *Consider model (2.1) and let β^{gS} be an estimate given by (2.2). We define two random variables: the number of all groups selected by gSLOPE (Rg) and the number of groups falsely discovered by gSLOPE (Vg), as*

$$Rg := \left|\left\{i : \ \|X_{I_i}\beta^{\text{gS}}_{I_i}\|_2 \neq 0\right\}\right|, \qquad Vg := \left|\left\{i : \ \|X_{I_i}\beta_{I_i}\|_2 = 0, \ \|X_{I_i}\beta^{\text{gS}}_{I_i}\|_2 \neq 0\right\}\right|.$$

Definition 2.3. *We define the false discovery rate for groups (gFDR) as*

$$gFDR = gFDR\big(X, \beta, \sigma^2\big) := \mathbb{E}\left[\frac{Vg}{\max\{Rg, 1\}}\right]. \tag{2.9}$$

2.4 Control of gFDR when variables from different groups are orthogonal

Our goal is the identification of the regularizing sequence for gSLOPE such that gFDR can be controlled at any given level $q \in (0, 1)$. In this section we will provide such a sequence, which provably controls gFDR in case when variables in different groups are orthogonal to each other. In subsequent sections we will replace this condition with the weaker assumption of the stochastic independence of regressors in different groups. Before the statement of the main theorem on gFDR control, we will recall the definition of χ distribution and define a scaled χ distribution.

Definition 2.4. *We will say that a random variable X_1 has a χ distribution with l degrees of freedom, and write $X_1 \sim \chi_l$, when X_1 could be expressed as $X_1 = \sqrt{X_2}$, for X_2 having a χ^2 distribution with l degrees of freedom. We will say that a random variable X_1 has a scaled χ distribution with l degrees of freedom and scale $\mathcal{S}$, when X_1 could be expressed as $X_1 = \mathcal{S} \cdot X_2$, for X_2 having a χ distribution with l degrees of freedom. We will use the notation $X_1 \sim \mathcal{S}\chi_l$.*

Theorem 2.5 (gFDR control under orthogonal case)**.** *Consider model (2.1) with the design matrix X satisfying $X_{I_i}^{\mathsf{T}} X_{I_j} = 0$, for any $i \neq j$. Denote the number of zero coefficients in $[\![\beta]\!]_{I,X}$ by m_0 and let $w_1, \ldots, w_m$ be positive numbers. Moreover, define the sequence of regularizing parameters $\lambda^{\max} = (\lambda_1^{\max}, \ldots, \lambda_m^{\max})^{\mathsf{T}}$, with*

$$\lambda_i^{\max} := \max_{j=1,\ldots,m} \left\{ \frac{1}{w_j} F_{\chi_{l_j}}^{-1} \left(1 - \frac{q \cdot i}{m} \right) \right\}, \tag{2.10}$$

where $F_{\chi_{l_j}}$ is a cumulative distribution function of χ distribution with l_j degrees of freedom. Then any solution, β^{gS}, to problem gSLOPE (2.2) generates the same vector $[\![\beta^{\mathrm{gS}}]\!]_{I,X}$ and it holds

$$gFDR = \mathbb{E}\left[\frac{Vg}{\max\{Rg, 1\}} \right] \leq q \cdot \frac{m_0}{m}.$$

Proof. Consider the standardized version of the gSLOPE problem, given by (2.4). Since X is orthogonal at groups level, $\widetilde{X}$ in problem (2.4) is an orthogonal matrix, i.e. $\widetilde{X}^{\mathsf{T}}\widetilde{X} = \mathbf{I}_{\widetilde{p}}$. It is easy to show that this implies $\left\|y - \widetilde{X}b\right\|_2^2 = \left\|\widetilde{X}^{\mathsf{T}}y - b\right\|_2^2 + C$, where C does not depend on b (see Appendix D). Hence under orthogonal situation the optimization problem in (2.4) can be recast as

$$b^* := \underset{b}{\operatorname{argmin}} \quad \left\{ \frac{1}{2}\left\|\widetilde{y} - b\right\|_2^2 + \sigma J_\lambda\big(W[\![b]\!]_{\mathbb{I}}\big) \right\}, \tag{2.11}$$

where $\widetilde{y} := \widetilde{X}^{\mathsf{T}}y$ is a whitened version of y, which has the multivariate normal distribution $\mathcal{N}\big(\widetilde{\beta},\ \sigma^2\mathbf{I}_{\widetilde{p}}\big)$, with $\widetilde{\beta}_{\mathbb{I}_i} := R_i\beta_{I_i}$, $i = 1, \ldots, m$. As discussed in the previous section the problem (2.11) can be equivalently formulated as

$$\begin{cases} c^* = \operatorname{argmin}_c \left\{ \frac{1}{2}\sum_{i=1}^m \left(\|\widetilde{y}_{\mathbb{I}_i}\|_2 - w_i^{-1}c_i\right)^2 + J_{\sigma\lambda}(c) \right\} \\ b^*_{\mathbb{I}_i} = c_i^*\big(w_i\|\widetilde{y}_{\mathbb{I}_i}\|_2\big)^{-1}\widetilde{y}_{\mathbb{I}_i}, \quad i = 1, \ldots, m \end{cases}. \tag{2.12}$$

The above formulation yields the conclusion, that indices of groups estimated by gSLOPE as relevant coincide with the support of the solution to the SLOPE problem with the diagonal design matrix D such that $D_{ii} = w_i^{-1}$. After defining $\widetilde{\beta} \in \mathbb{R}^{\widetilde{p}}$ by conditions $\widetilde{\beta}_{\mathbb{I}_i} := R_i\beta_{I_i}$, $i = 1, \ldots, m$, we also have $\widetilde{y} \sim \mathcal{N}\big(\widetilde{\beta},\ \sigma^2\mathbf{I}_{\widetilde{p}}\big)$.

Now, we define random variables $R := \left|\left\{i :\ c_i^* \neq 0\right\}\right|$ and $V := \left|\left\{i :\ \|\widetilde{\beta}_{\mathbb{I}_i}\|_2 = 0, \quad c_i^* \neq 0\right\}\right|$. Clearly, then $Rg = R$ and $Vg = V$. Consequently, it is enough to show that

$$\mathbb{E}\left[\frac{V}{\max\{R, 1\}}\right] \leq q \cdot \frac{m_0}{m}.$$

Without loss of generality we can assume that groups $I_1, \ldots, I_{m_0}$ are truly irrelevant, which gives $\|\widetilde{\beta}_{\mathbb{I}_1}\|_2 = \ldots = \|\widetilde{\beta}_{\mathbb{I}_{m_0}}\|_2 = 0$ and $\|\widetilde{\beta}_{\mathbb{I}_j}\|_2 > 0$ for $j > m_0$. Suppose now that r, i are some fixed indices from $\{1, \ldots, m\}$. From definition of $\lambda_r^{\max}$

$$\lambda_r^{\max} \geq \frac{1}{w_i}F_{\chi_{l_i}}^{-1}\left(1 - \frac{qr}{m}\right) \implies 1 - F_{\chi_{l_i}}\left(\lambda_r^{\max}w_i\right) \leq \frac{qr}{m}. \tag{2.13}$$

Now, let us assume that $i \leq m_0$. Since $\sigma^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 \sim \chi_{l_i}$ we have

$$\mathbb{P}\left(w_i^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 \geq \sigma\lambda_r^{\max}\right) = \mathbb{P}\left(\sigma^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 \geq \lambda_r^{\max} w_i\right) = 1 - F_{\chi_{l_i}}\left(\lambda_r^{\max} w_i\right) \leq \frac{qr}{m}. \quad (2.14)$$

Denote by $\widetilde{R}^i$ the number of nonzero coefficients in SLOPE estimate (2.12) after eliminating i^{th} group of explanatory variables. Thanks to lemmas E.6 and E.7 in the Appendix, we immediately get

$$\left\{[\![\widetilde{y}]\!]_{\mathbb{I}} :\ c_i^* \neq 0 \text{ and } R = r\right\} \subset \left\{[\![\widetilde{y}]\!]_{\mathbb{I}} :\ w_i^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 > \sigma\lambda_r^{\max} \text{ and } \widetilde{R}^i = r-1\right\}, \quad (2.15)$$

which together with (2.14) raises

$$\begin{aligned}
\mathbb{P}(c_i^* \neq 0 \text{ and } R = r) \ &\leq\ \mathbb{P}\left(w_i^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 > \sigma\lambda_r^{\max} \text{ and } \widetilde{R}^i = r-1\right) \\
&=\ \mathbb{P}\left(w_i^{-1}\|\widetilde{y}_{\mathbb{I}_i}\|_2 > \sigma\lambda_r^{\max}\right)\mathbb{P}\left(\widetilde{R}^i = r-1\right) \\
&\leq\ \frac{qr}{m}\mathbb{P}\left(\widetilde{R}^i = r-1\right),
\end{aligned} \quad (2.16)$$

where the equality follows from the independence between $\|\widetilde{y}_{\mathbb{I}_i}\|_2$ and $\widetilde{R}^i$. Therefore

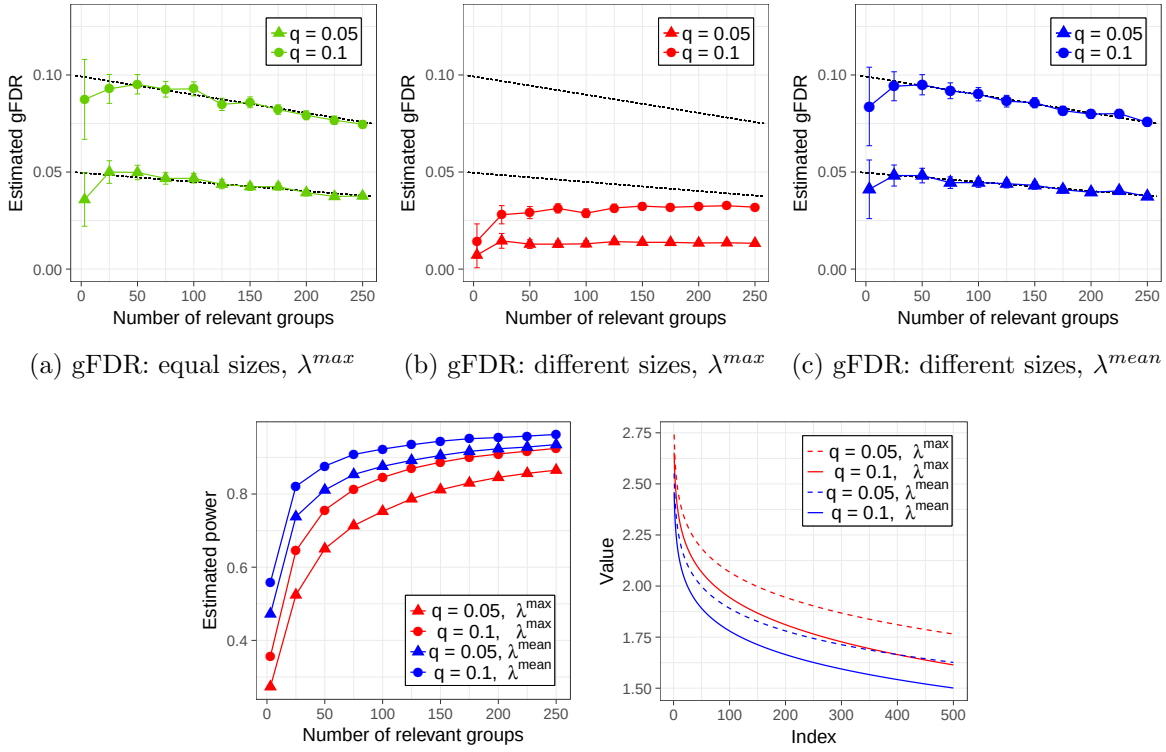
$$\begin{aligned}
&\mathbb{E}\left[\frac{V}{\max\{R,1\}}\right] = \sum_{r=1}^{m}\mathbb{E}\left[\frac{V}{r}\mathbb{1}_{\{R=r\}}\right] = \sum_{r=1}^{m}\frac{1}{r}\mathbb{E}\left[\sum_{i=1}^{m_0}\mathbb{1}_{\{c_i^*\neq 0\}}\mathbb{1}_{\{R=r\}}\right] = \\
&\sum_{r=1}^{m}\frac{1}{r}\sum_{i=1}^{m_0}\mathbb{P}\left(c_i^* \neq 0 \text{ and } R = r\right) \ \leq\ \sum_{i=1}^{m_0}\frac{q}{m}\sum_{r=1}^{m}\mathbb{P}\big(\widetilde{R}^i = r-1\big) = \frac{qm_0}{m},
\end{aligned} \quad (2.17)$$

which finishes the proof. ■

Figure 2 illustrates the performance of gSLOPE under the design matrix $X = \mathbf{I}_p$ (hence l_i, the rank of group i, coincides with its size), with $p = 5000$. In Figure 2 (a) all groups are of the same size $l = 5$, while in Figures 2 (b) - (d) the explanatory variables are clustered into $m = 1000$ groups of sizes from the set $\{3, 4, 5, 6, 7\}$; 200 groups of each size. Each coefficient of β_{I_i}, in a truly relevant group i, was generated independently from a $U[0.1, 1.1]$ distribution and then β_{I_i} was scaled such that $\left([\![\beta]\!]_{I,X}\right)_i = a\sqrt{l_i}$. Parameter a was selected to satisfy the condition

$$a\sum_{i=1}^{m}\sqrt{l_i} = \sum_{i=1}^{m}\sqrt{4\ln(m)/(1-m^{-2/l_i}) - l_i}\ ,$$

which, according to the calculations presented in the Appendix H, yields signals comparable to the maximal noise. Such signals can be detected with moderate power, which allows for a meaningful comparison between different methods.



(a) gFDR: equal sizes, λ^{max} (b) gFDR: different sizes, λ^{max} (c) gFDR: different sizes, λ^{mean}

(d) power: different sizes (e) lambdas: different sizes

Figure 2: Orthogonal situation with $n = p = 5000$ and $m = 1000$. In (a) all groups are of the same size $l = 5$, while in (b)-(d) there are 200 groups of each of sizes $l_i \in \{3, 4, 5, 6, 7\}$. In (a) and (b) gSLOPE works with the regularizing sequence λ^{max}, while in (c) and (d) λ^{mean} is used. First 500 elements of different λ sequences are shown in (e). For each target gFDR level and true support size, 300 iterations were performed. Bars correspond to $\pm 2SE$. Black straight lines represent the "nominal" gFDR level $q \cdot \big((m-k)/m\big)$, for k being true support size. Weights are defined as $w_i := \sqrt{l_i}$.

Figure 2 (a) illustrates that the sequence $\lambda^{\max}$ keeps gFDR very close to the "nominal" level when groups are of the same size. However, Figure 2 (b) shows that for groups of different size $\lambda^{\max}$ is rather conservative, i.e. the achieved gFDR is significantly lower than assumed. This suggests that the shrinkage (dictated by λ) could be decreased, such that the method gets more power and still achieves the gFDR below the assumed level. Returning to the proof of Theorem 2.5, we can see that for each

$i \in \{1, \ldots, m\}$ we have

$$1 - F_{\chi_{l_i}}(\lambda_r^{\max} w_i) \leq \frac{qr}{m}, \tag{2.18}$$

with equality holding only for i being the index of the maximum in (2.10). In the result the inequality in (2.17) is usually strict and the true gFDR might be substantially smaller than the nominal level. The natural relaxation of (2.18) is to require only that

$$\sum_{i=1}^{m} \left(1 - F_{w_i^{-1}\chi_{l_i}}(\lambda_r)\right) \leq qr. \tag{2.19}$$

Replacing the inequality in (2.19) by equality yields the strategy of choosing the relaxed λ sequence

$$\lambda_r^{\text{mean}} := \overline{F}^{-1}\left(1 - \frac{qr}{m}\right) \quad \text{for} \quad \overline{F}(x) := \frac{1}{m}\sum_{i=1}^{m} F_{w_i^{-1}\chi_{l_i}}(x), \quad r \in \{1, \ldots, m\}, \tag{2.20}$$

where $F_{w_i^{-1}\chi_{l_i}}$ is the cumulative distribution function of scaled chi distribution with l_i degrees of freedom and scale $\mathcal{S} = w_i^{-1}$. In Figure 2 (c) we present estimated gFDR, for tuning parameters given by (2.20). The results suggest that with a relaxed version of tuning parameters, we can still achieve the "average" gFDR control, where the "average" is with respect to the uniform distribution over all possible signal placements. As shown in Figure 2 (d), application of λ^{mean} allows to achieve a substantially larger power than the one provided by $\lambda^{\max}$. Such a strategy could be especially important in situations where differences between the smallest and the largest quantiles (among distributions $w_i^{-1}\chi_{l_i}$) are relatively large and all groups have the same prior probability of being relevant.

2.5 The accuracy of estimation

Up until this point, we have only considered the testing properties of gSLOPE. Though originally proposed to control the FDR, surprisingly, SLOPE enjoys appealing estimation properties as well, (see e.g Su and Candès, 2016; Bellec et al., 2016b,a). It thus would be desirable to extend this link between testing and estimation for gSLOPE.

In measuring the deviation of an estimator from the ground truth β, as earlier, we focus on the group instead of an individual level. Accordingly, here we aim to estimate parts of variance of Y explained by every group, which are contained in the vector $[\![\beta]\!]_{X,I} := \left(\|X_{I_1}\beta_{I_1}\|_2, \ldots, \|X_{I_m}\beta_{I_m}\|_2\right)^{\mathsf{T}}$ or $[\![\widetilde{\beta}]\!]_{\mathbb{I}} := \left(\|\widetilde{\beta}_{\mathbb{I}_1}\|_2, \ldots, \|\widetilde{\beta}_{\mathbb{I}_m}\|_2\right)^{\mathsf{T}}$, equivalently. For illustration purpose, we employ the setting described as follows. Imagine that we have a sequence of problems with the number of groups m growing to infinity: the design X is orthonormal at groups level; ranks of submatrices X_{I_i}, l_i, are bounded, that is, $\max l_i \leq l$ for some constant integer l; denoting by $k \geq 1$ the sparsity level (that is, the number of relevant groups), we assume the asymptotics $k/m \to 0$. Now we state our minimax theorem, where we write $a \sim b$ if $a/b \to 1$ in the asymptotic limit, and $\|[\![\beta]\!]_{I,X}\|_0$ denotes the number of nonzero entries of $[\![\beta]\!]_{I,X}$. The proof makes use of the same techniques for proving Theorem 1.1 in (Su and Candès, 2016) and is deferred to the Appendix.

Theorem 2.6. *Fix any constant $q \in (0,1)$, let $w_i = 1$ and $\lambda_i = F^{-1}_{\chi_l}(1 - qi/m)$ for $i = 1, \ldots, m$. Under the preceding conditions $m \to \infty$ and $k/m \to 0$, gSLOPE is asymptotically minimax over the nearly black object $\left\{\beta : \left\|[\![\beta]\!]_{I,X}\right\|_0 \leq k\right\}$, i.e.,*

$$\sup_{\|[\![\beta]\!]_{I,X}\|_0 \leq k} \mathbb{E}\left(\left\|[\![\beta^{\mathrm{gS}}]\!]_{I,X} - [\![\beta]\!]_{I,X}\right\|_2^2\right) \sim \inf_{\widehat{\beta}} \sup_{\|[\![\beta]\!]_{I,X}\|_0 \leq k} \mathbb{E}\left(\left\|[\![\widehat{\beta}]\!]_{I,X} - [\![\beta]\!]_{I,X}\right\|_2^2\right),$$

where the infimum is taken over all measurable estimators $\widehat{\beta}(y, X)$.

Notably, in this theorem the choice of λ_i does not assume the knowledge of sparsity level. Or putting it differently, in stark contrast to gLASSO, gSLOPE is adaptive to a range of sparsity in achieving the exact minimaxity. Combining Theorems 2.5 and 2.6, we see the remarkable link between FDR control and minimax estimation also applies to gSLOPE (Abramovich et al., 2006; Su and Candès, 2016). While it is out of the scope of this paper, it is of great interest to extend this minimax result to general design matrices.

2.6 The impact of chosen weights

In this subsection we will discuss the influence of chosen weights, $\{w_i\}_{i=1}^m$, on results. Let $I = \{I_1, \ldots, I_m\}$ be a given partition into groups and $l_1, \ldots, l_m$ be ranks of submatrices X_{I_i}. Assume the orthogonality at group level, i.e., that it holds $X_{I_i}^{\mathsf{T}} X_{I_j} = 0$, for $i \neq j$, and suppose that $\sigma = 1$. The support of $[\![\beta]\!]_{I,X}$ coincides with the support of vector c^* defined in (2.12), namely

$$c^* = \underset{c}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| [\![\widetilde{y}]\!]_{\mathbb{I}} - W^{-1} c \right\|_2^2 + J_\lambda(c) \right\}, \tag{2.21}$$

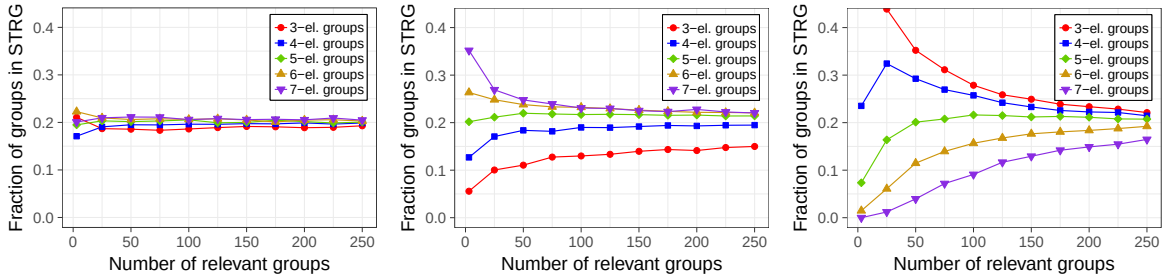
where W^{-1} is a diagonal matrix with positive numbers $w_1^{-1}, \ldots, w_m^{-1}$ on the diagonal. Suppose now, that c^* has exactly r nonzero coefficients. From Corollary E.4 in the Appendix E, these indices are given by $\{\pi(1), \ldots, \pi(r)\}$, where π is permutation which orders $W^{-1}[\![\widetilde{y}]\!]_{\mathbb{I}}$. Hence, the order of realizations $\left\{w_i^{-1} \|\widetilde{y}_{\mathbb{I}_i}\|_2\right\}_{i=1}^m$ decides about the subset of groups labeled by gSLOPE as relevant. Suppose that groups I_i and I_j are truly relevant, i.e., $\|\widetilde{\beta}_{\mathbb{I}_i}\|_2 > 0$ and $\|\widetilde{\beta}_{\mathbb{I}_j}\|_2 > 0$. The distributions of $\|\widetilde{y}_{\mathbb{I}_i}\|_2$ and $\|\widetilde{y}_{\mathbb{I}_j}\|_2$ are noncentral χ distributions, with l_i and l_j degrees of freedom, and the noncentrality parameters equal to $\|\widetilde{\beta}_{\mathbb{I}_i}\|_2$ and $\|\widetilde{\beta}_{\mathbb{I}_j}\|_2$, respectively. Now, the expected value of the noncentral χ distribution could be well approximated by the square root of the expected value of the noncentral χ^2 distribution, which gives

$$\mathbb{E}\big(w_i^{-1} \|\widetilde{y}_{\mathbb{I}_i}\|_2\big) \approx w_i^{-1} \sqrt{\mathbb{E}\big(\|\widetilde{y}_{\mathbb{I}_i}\|_2^2\big)} = w_i^{-1} \sqrt{l_i + \|\widetilde{\beta}_{\mathbb{I}_i}\|_2^2}.$$

Therefore, roughly speaking, truly relevant groups I_i and I_j are treated as comparable, when it occurs $l_i/w_i^2 + \|\widetilde{\beta}_{\mathbb{I}_i}\|_2^2/w_i^2 \approx l_j/w_j^2 + \|\widetilde{\beta}_{\mathbb{I}_j}\|_2^2/w_j^2$. This gives us the intuition about the behavior of gSLOPE with the choice $w_i = \sqrt{l_i}$ for each i. Firstly, gSLOPE treats all irrelevant groups as comparable, i.e. the size of the group has a relatively small influence on it being selected as a false discovery. Secondly, gSLOPE treats two truly relevant groups as comparable, if groups effect sizes satisfy

the condition $\left(\llbracket\beta\rrbracket_{I,X}\right)_i/\left(\llbracket\beta\rrbracket_{I,X}\right)_j \approx \sqrt{l_i}/\sqrt{l_j}$. The derived condition could be recast as $\|X_{I_i}\beta_{I_i}\|_2^2/l_i \approx \|X_{I_j}\beta_{I_j}\|_2^2/l_j$. This gives a nice interpretation: with the choice $w_i := \sqrt{l_i}$, gSLOPE treats two groups as comparable, when these groups have similar squared effect group sizes per coefficient. One possible idealistic situation, when such a property occurs, is when all β_i's in truly relevant groups are comparable.

In Figure 3 we see that when the condition $\left(\llbracket\beta\rrbracket_{I,X}\right)_i/\left(\llbracket\beta\rrbracket_{I,X}\right)_j = \sqrt{l_i}/\sqrt{l_j}$ is met, the fractions of groups with different sizes in the selected truly relevant groups (STRG) are approximately equal. To investigate the impact of selected weights on the set of discovered groups, we performed simulations with different settings, namely we used $w_i = 1$ and $w_i = l_i$ (without changing other parameters). With the first choice, larger groups are penalized less than before, while the second choice yields the opposite situation. This is reflected in the proportion of each groups in STRG (Figure 3). The



(a) STRG for $w_i := \sqrt{l_i}$ (b) STRG for $w_i := 1$ (c) STRG for $w_i := l_i$

Figure 3: Fraction of each group sizes in selected truly relevant groups (STRG). Beyond the weights, this simulation was conducted with the same setting as in experiments summarized in Figure 1 for λ^{mean}. *In particular, for truly relevant groups* i *and* j, *it occurs* $\left(\llbracket\beta\rrbracket_{I,X}\right)_i/\left(\llbracket\beta\rrbracket_{I,X}\right)_j = \sqrt{l_i}/\sqrt{l_j}$. *Target gFDR level was fixed as* 0.05.

values of gFDR are very similar under all choices of weights.

2.7 Independent groups and unknown σ

The assumption that variables in different groups are orthogonal to each other can be satisfied only in rare situations of specifically designed experiments. However, in a va-

riety of applications one can assume that variables in different groups are independent. Such a situation occurs for example in the context of identifying influential genes using distant genetic markers, whose genotypes can be considered as stochastically independent. In this case a group can be formed by clustering dummy variables corresponding to different genotypes of a given marker. Though the difference between stochastic independence and algebraic orthogonality seems rather small, it turns out that small sample correlations between independent regressors together with the shrinkage of regression coefficients lead to magnifying the effective noise and require the adjustment of the tuning sequence λ (see Su et al., 2015, for discussion of this phenomenon in the context of LASSO). Concerning regular SLOPE, this problem was addressed by heuristic modification of λ, proposed in (Bogdan et al., 2013, 2015). This modified sequence was calculated based upon the assumption that explanatory variables are randomly sampled from the Gaussian distribution. However, simulation results from (Bogdan et al., 2015) illustrate that it controls FDR also in case when the columns of the design matrix correspond to additive effects of independent SNPs and the number of causal genes is moderately small.

To derive the similar heuristic adjustment for the group SLOPE we will at first confine ourselves to the case $\sigma = 1$, $l_1 = \ldots = l_m := l$, $w_1 = \ldots = w_m := w$.

The first step in our derivation relies on specifying the optimality conditions for the standardized version of group SLOPE provided in (2.4), under which $X_{I_i}^{\mathsf{T}} X_{I_i} = \mathbf{I}_l$, for all $i \in \{1, \ldots, m\}$.

Theorem 2.7 (Optimality conditions)**.** *Let X be the standardized design matrix satisfying $X_{I_i}^{\mathsf{T}} X_{I_i} = \mathbf{I}_l$ for each i and let $\hat{\beta}$ be the solution to the gSLOPE problem. Let us order the groups such that $\|\hat{\beta}_{I_1}\|_2 > \ldots > \|\hat{\beta}_{I_s}\|_2 > 0$ and $\|\hat{\beta}_{I_i}\|_2 = 0$ for $i > s$ and consider the partition of I into $I^S := \{I_1, \ldots, I_s\}$ and $I^C := \{I_{s+1}, \ldots, I_m\}$. Moreover, let us define $v_{I_i} := X_{I_i}^{\mathsf{T}} \left(y - X_{\setminus I_i} \hat{\beta}_{\setminus I_i} \right)$, where $X_{\setminus I_i}$ is a matrix X without columns from*

I_i. Then the following two sets of conditions are met

$$\begin{cases} X_{I_i}^{\mathsf{T}}(y - X\hat{\beta}) = w\lambda_i \frac{\hat{\beta}_{I_i}}{\|\hat{\beta}_{I_i}\|_2}, \ i \le s \\ [\![X^{\mathsf{T}}(y - X\hat{\beta})]\!]_{I^c} \in C_{w\lambda^c} \end{cases} \text{and} \quad \begin{cases} \big| \|v_{I_i}\|_2 - w\lambda_i \big| = \|\hat{\beta}_{I_i}\|_2, \ i \le s \\ [\![v]\!]_{I^c} \in C_{w\lambda^c} \end{cases}, \quad (2.22)$$

where $v := (v_{I_1}^{\mathsf{T}}, \ldots, v_{I_m}^{\mathsf{T}})^{\mathsf{T}}$, $\lambda^c = (\lambda_{s+1}, \ldots, \lambda_m)$ *and* $C_{w\lambda^c} := \Big\{ x \in \mathbb{R}^{m-s} : \ \sum_{i=1}^{k} |x|_{(i)} \le \sum_{i=1}^{k} w\lambda_i^c, \ k = 1, \ldots, m-s \Big\}$ *is the unit ball of the dual norm to* $J_{w\lambda^c}$.

Proof. The proof of Theorem 2.7 is provided in Appendix G. ■

The task now is to select λ_i's such that the condition $[\![v]\!]_{I^C} \in C_{w\lambda^c}$ regulates the rate of false discoveries. Let us at first observe that

$$v_{I_i} = X_{I_i}^{\mathsf{T}} \Big(X\beta - X_{\setminus I_i}\hat{\beta}_{\setminus I_i} + z \Big) = \beta_{I_i} + X_{I_i}^{\mathsf{T}} X_{\setminus I_i}(\beta_{\setminus I_i} - \hat{\beta}_{\setminus I_i}) + X_{I_i}^{\mathsf{T}} z. \quad (2.23)$$

Note that under the orthogonal design the last expression reduces to $X_{I_i}^{\mathsf{T}} z$ for $i > s$ and has χ distribution with l degrees of freedom. This fact was used in subsection 2.4 to define the sequence λ. In the considered near-orthogonal situation, the term $X_{I_i}^{\mathsf{T}} X_{\setminus I_i}(\beta_{\setminus I_i} - \hat{\beta}_{\setminus I_i})$ does not vanish and creates an additional "noise", which needs to be taken into account when designing the λ sequence. To approximate the distribution of v_{I_i} under the assumption of independence between different groups we will use the following simplifying assumptions. To estimate the distribution of v_{I_i} we will first simplify the situation by assuming that true and estimated signals define the same set of relevant groups, with indices from the set $\{1, \ldots, s\}$, and that the signal strength is sufficiently large to assume that $\frac{\hat{\beta}_{I_i}}{\|\hat{\beta}_{I_i}\|_2}$ can be well approximated by $\frac{\beta_{I_i}}{\|\beta_{I_i}\|_2}$ for $i \le s$. After defining $I_S := \bigcup_{i \le s} I_i$, from the left set of conditions in (2.22) we get that

$$X_{I_S}^{\mathsf{T}}(X_{I_S}\beta_{I_S} - X_{I_S}\hat{\beta}_{I_S}) + X_{I_S}^{\mathsf{T}} z \approx w \underbrace{\big(\lambda_1 \beta_{I_1}^{\mathsf{T}} / \|\beta_{I_1}\|_2, \ldots, \lambda_s \beta_{I_s}^{\mathsf{T}} / \|\beta_{I_s}\|_2 \big)^{\mathsf{T}}}_{H_{\lambda,\beta}}, \quad (2.24)$$

which gives $X_{I_i}^{\mathsf{T}} X_{I_S}(\beta_{I_S} - \hat{\beta}_{I_S}) \approx X_{I_i}^{\mathsf{T}} X_{I_S}(X_{I_S}^{\mathsf{T}} X_{I_S})^{-1}(wH_{\lambda,\beta} - X_{I_S}^{\mathsf{T}} z)$. Finally, combining the last expression with (2.23) let us to assume $v_{I_i} \approx \hat{v}_{I_i}$, where

$$\hat{v}_{I_i} := X_{I_i}^{\mathsf{T}} X_{I_S} (X_{I_S}^{\mathsf{T}} X_{I_S})^{-1} \Big(w H_{\lambda,\beta} - X_{I_S}^{\mathsf{T}} z \Big) + X_{I_i}^{\mathsf{T}} z. \tag{2.25}$$

Now, we will assume that the distribution of $\hat{v}_{I_i}$ can be well approximated by assuming that the individual entries of X come from the normal distribution $\mathcal{N}\big(0, \frac{1}{n}\big)$. This assumption can be justified if the distribution of the individual entries of X is sufficiently regular and n is substantially larger than lm_0. The following Theorem 2.8 provides the expected value and the covariance matrix of the random vector $\hat{v}_{I_i}$ for $i > s$ under the assumption of normality.

Theorem 2.8. *Assume that the entries of the design matrix X are independently drawn from $\mathcal{N}(0, 1/n)$ distribution. Then for each $i > s$ the expected value of $\hat{v}_{I_i}$ is equal to 0 and the covariance matrix of this random vector is given by*

$$Cov(\hat{v}_{I_i}) = \left(\frac{n - ls}{n} + w^2 \frac{\|\lambda^S\|_2^2}{n - ls - 1} \right) \mathbf{I}_l, \quad \textit{where} \quad \lambda^S := (\lambda_1, \ldots, \lambda_s)^{\mathsf{T}}. \tag{2.26}$$

Proof. The proof of Theorem 2.8 is provided in Appendix G. ■

Now, if n is large enough with respect to sl, then by the Central Limit Theorem the distribution of v_{I_i} can be approximated by the multivariate normal distribution and the distribution of $\|v_{I_i}\|_2$ by the scaled χ distribution with l degrees of freedom and a scale parameter $\mathcal{S} = \sqrt{\frac{n-ls}{n} + \frac{w^2 \|\lambda_S\|_2^2}{n-sl-1}}$. Now, analogously to the orthogonal situation, lambdas could be defined as $\lambda_i := \frac{1}{w_i} F^{-1}_{\mathcal{S}\chi_l}\left(1 - \frac{q \cdot i}{m}\right) = \frac{\mathcal{S}}{w_i} F^{-1}_{\chi_l}\left(1 - \frac{q \cdot i}{m}\right)$. Since s is unknown, we will apply the strategy used in (Bogdan et al., 2013): define λ_1 as in orthogonal case and for $i \geq 2$ define λ_i by incorporating the scale parameter corresponding to the sparsity $s = i - 1$. This yields the following procedure.

Procedure 2 Selecting lambdas under the assumption of independence: equal groups sizes

input: $q \in (0,1),\ \ w > 0,\ \ p,\ n,\ m,\ l \in \mathbb{N}$
$\lambda_1 := \frac{1}{w} F^{-1}_{\chi_l}\left(1 - \frac{q}{m}\right)$;
For $i \in \{2, \ldots, m\}$:
$\lambda^S := (\lambda_1, \ldots, \lambda_{i-1})^{\mathsf{T}}$;
$\mathcal{S} := \sqrt{\frac{n - l(i-1)}{n} + \frac{w^2 \|\lambda^S\|_2^2}{n - l(i-1) - 1}}$;
$\lambda_i^* := \frac{\mathcal{S}}{w} F^{-1}_{\chi_l}\left(1 - \frac{q \cdot i}{m}\right)$;
if $\lambda_i^* \leq \lambda_{i-1}$, then put $\lambda_i := \lambda_i^*$. Otherwise, stop the procedure and put $\lambda_j := \lambda_{i-1}$ for $j \geq i$;
end for

Consider now the Gaussian design with arbitrary group sizes and a sequence of positive weights $w_1, \ldots, w_m$. One possible approach is to construct consecutive λ_i by taking the largest scaled quantiles among all distributions, i.e. as $\max_{j=1,\ldots,m} \left\{ \frac{\mathcal{S}_j}{w_j} F^{-1}_{\chi_{l_j}}\left(1 - \frac{q \cdot i}{m}\right) \right\}$, with the scale parameter $\mathcal{S}_j$ adjusted to l_j (the conservative strategy). In this article, however, we will stick to the more liberal strategy based on λ^{mean}, which leads to the modified sequence of tuning parameters presented in Procedure 3.

Procedure 3 Sequence of tuning parameters for independent groups

input: $q \in (0,1),\ \ w_1, \ldots, w_m > 0,\ \ p,\ n,\ m,\ l_1, \ldots, l_m \in \mathbb{N}$
$\lambda_i := \overline{F}^{-1}\left(1 - \frac{q \cdot i}{m}\right), \quad \text{for} \quad \overline{F}(x) := \frac{1}{m} \sum_{i=1}^m F_{w_i^{-1} \chi_{l_i}}(x)$;
for $i \in \{2, \ldots, m\}$:
$\lambda^S := (\lambda_1, \ldots, \lambda_{i-1})^{\mathsf{T}}$;
$\mathcal{S}_j := \sqrt{\frac{n - l_j(i-1)}{n} + \frac{w_j^2 \|\lambda^S\|_2^2}{n - l_j(i-1) - 1}}, \qquad \text{for } j \in \{1, \ldots, m\}$;
$\lambda_i^* := \overline{F}_{\mathcal{S}}^{-1}\left(1 - \frac{qi}{m}\right), \quad \text{for} \quad \overline{F}_{\mathcal{S}}(x) := \frac{1}{m} \sum_{j=1}^m F_{\mathcal{S}_j w_j^{-1} \chi_{l_j}}(x)$;
if $\lambda_i^* \leq \lambda_{i-1}$, then put $\lambda_i := \lambda_i^*$. Otherwise, stop the procedure and put $\lambda_j := \lambda_{i-1}$ for $j \geq i$;
end for

Up until this moment, we have used σ in gSLOPE optimization problem, assuming that this parameter is known. However, in many applications σ is unknown and its estimation is an important issue. When $n > p$, the standard procedure is to use the unbiased estimator of σ^2, $\hat{\sigma}^2_{\text{OLS}}$, given by

$$\hat{\sigma}^2_{\text{OLS}} := \left(y - X\beta^{\text{OLS}}\right)^{\mathsf{T}} \left(y - X\beta^{\text{OLS}}\right)/(n-p), \text{ for } \beta^{\text{OLS}} := (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y. \tag{2.27}$$

For the target situation, with p much larger than n, such an estimator can not be used. To estimate σ we will therefore apply the procedure which was dedicated for this

purpose in (Bogdan et al., 2015) in the context of SLOPE. Below we present algorithm adjusted to gSLOPE (Procedure 4). The idea standing behind the procedure is simple.

Procedure 4 gSLOPE with estimation of σ

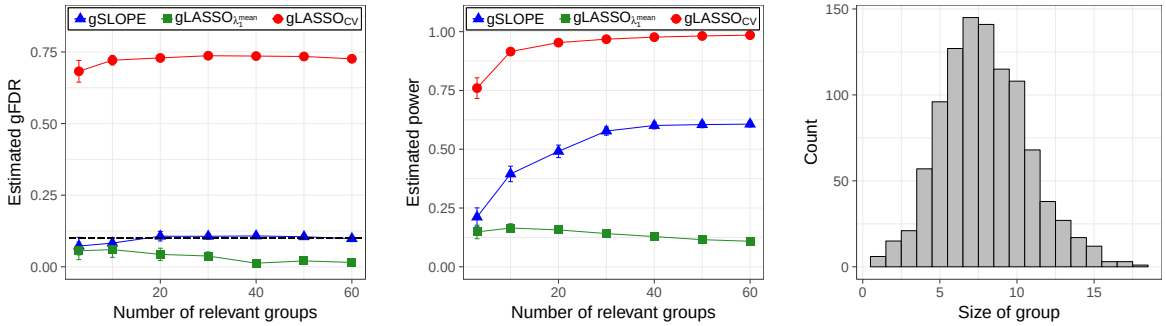
input: y, X and λ (defined for some fixed q)
initialize: $S_+ = \emptyset$;
repeat
  $S = S_+$;
  compute RSS obtained by regressing y onto variables in S;
  set $\hat{\sigma}^2 = RSS/(n - |S| - 1)$;
  compute the solution β^{gS} to gSLOPE with parameters $\hat{\sigma}$ and sequence λ;
  set $S_+ = \mathrm{supp}(\beta^{\mathrm{gS}})$;
until $S_+ = S$

The gSLOPE property of producing sparse estimators is used, and in each iteration columns in design matrix are first restricted to support of β^{gS}, so that the number of rows exceeds the number of columns and (2.27) can be used. Algorithm terminates when gSLOPE finds the same subset of relevant variables as in the preceding iteration.

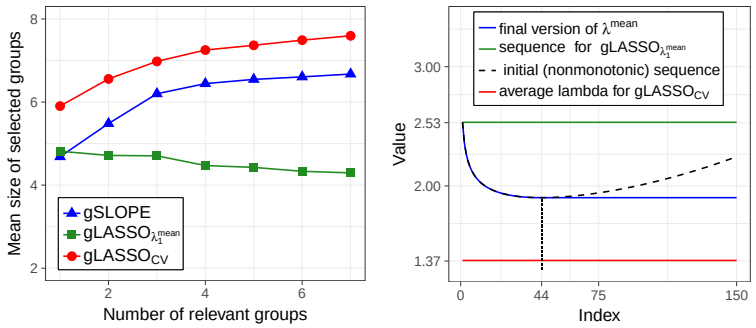
To investigate the performance of gSLOPE under the Gaussian design and various group sizes, we performed simulations with 1000 groups. Their sizes were drawn from the binomial distribution, $Bin(1000; 0.008)$, so as the expected value of the group size was equal to 8 (Figure 4 (c)). As a result, we obtained 7917 variables, divided into 1000 groups (the same division was used in all iterations and scenarios). For each sparsity level and the gFDR level 0.1, and each iteration we generated entries of the design matrix using $\mathcal{N}\big(0, \frac{1}{n}\big)$ distribution, then X was standardized and the values of response variable were generated according to model (2.1) with $\sigma = 1$ and signals generated as in simulations for Figure 2. To identify relevant groups based on the simulated data we have used the iterative version of gSLOPE, with σ estimation (Procedure 4) and lambdas given by Procedure 3. We performed 200 repetitions for each scenario, n was fixed as 5000. Results are represented in Figure 4 and show that our procedure allows to control gFDR at the assumed level.

Additionally, Figure 4 compares gSLOPE to gLASSO with two choices of the smoothing parameter λ. Firstly, we used $\lambda = \lambda_1^{\mathrm{mean}}$, which allows control of the gFDR

under the total null hypothesis. Secondly, for each of the iterations we chose λ based on leave-one-out cross-validation. It turns out that the first of these choices becomes rather conservative when the number of truly relevant groups increases. Then gLASSO has a smaller FDR but also a much smaller power than SLOPE (by a factor of three for $k = 60$). Cross-validation works in the opposite way - it yields a large power but also results in a huge proportion of false discoveries, which in our simulations systematically exceeds 60%.



(a) gFDR (b) Power (c) Histogram of group sizes



(d) Mean size of group (e) Lambdas sequences

Figure 4: Results for the example with independent regressors and various group sizes: $m = 1000$, $p = 7917$ *and* $n = 5000$. *Bars correspond to* $\pm 2SE$. *Entries of design matrix were drawn from* $\mathcal{N}(0, 1/n)$ *distribution and truly relevant signal,* i, *was generated such as* $\|X_{I_i}\beta_{I_i}\|_2 = \frac{1}{m}\sum_{i=1}^{m} B(m, l_i)$, *where* $B(m, l)$ *is defined in* (H.4).

Table 1: Coding for explanatory variables

	Genotype aa	Genotype aA	Genotype AA
additive dummy variable $\widetilde{X}$	2	1	0
dominance dummy variable $\widetilde{Z}$	0	1	0

2.8 Simulations under Genome-Wide Association Studies

To test the performance of gSLOPE in the context of Genome-Wide Association Studies (GWAS) we have used the North Finland Birth Cohort (NFBC1966) dataset, available in dbGaP with accession number phs000276.v2.p1 (`http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000276.v2.p1`) and described in detail in (Sabatti et al., 2009). The raw data contains $364,590$ markers for $5,402$ subjects. To obtain roughly independent SNPs this data set was initially screened using the *clump* procedure in the PLINK software (Purcell et al., 2007; Purcell, 2009) and additional screening in R such that in the final data set the maximal correlation between any pair of SNPs does not exceed $\sqrt{0.1} = 0.316$. The reduced data set contains $p = 26,315$ SNPs. The details of the screening procedure are provided in Appendix I.

The explanatory variables for our genetic model were defined in Table 1, where a denotes the less frequent (variant) allele. In case when population frequencies of both alleles are the same, variables $\widetilde{X}$ and $\widetilde{Z}$ are uncorrelated. In other cases correlations between these variables is different from zero and can be very strong for rare genetic variants. Since each SNP is described by two dummy variables, the full design matrix $[\widetilde{X}\ \widetilde{Z}]$ contains $52,630$ potential regressors. This matrix was then centered and standardized, so the columns of the final matrix $[X\ Z]$ have zero mean and unit norm.

The trait values are simulated according to two scenarios. In Scenario 1 we simulate from an additive model, where each of the causal SNPs influences the trait only through the additive dummy variable in matrix X,

$$y = X\beta_X + \varepsilon. \tag{2.28}$$

Here $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$, the number of 'causal' SNPs k varies between 1 and 80 and each causal SNP has an additive effect (non-zero components of β_X) equal to 5 or -5, with $P(\beta_{Xi} = 5) = P(\beta_{Xi} = -5) = 0.5$. In each of 100 iterations of our experiment causal SNPs were randomly selected from the full set of $26,315$ SNPs.

Since SNPs were selected in such a way that they are only weakly correlated, the identification of "causal" mutations based on the additive model (2.28) can be done with regular SLOPE, as it was demonstrated in (Bogdan et al., 2015). However, the additive model (2.28) implicitly assumes that for each of the SNPs the expected value of the trait for the heterozygote aA is the average of expected trait values for both homozygotes aa and AA. This idealistic assumption is usually not satisfied and many of the SNPs exhibit some dominance effects. To investigate the performance of model selection criteria in the presence of the dominance effects, we simulated data according to Scenario 2;

$$y = [X \ Z] \begin{bmatrix} \beta_X \\ \beta_Z \end{bmatrix} + \varepsilon \tag{2.29}$$

which differs from Scenario 1 by adding dominance effects (non-zero components of β_Z), which for each of k selected SNPs are sampled from the uniform distribution on$[-5, -3] \cup [3, 5]$. Now, the influence of i-th SNP on the trait is described by the vector $\boldsymbol{\beta}_i = (\beta_{Xi}, \beta_{Zi})$, containing its additive and dominance effects, which sets the stage for the application of the gSLOPE.

The data simulated according to Scenario 1 and Scenario 2 were analyzed using three different approaches:

A.1 gSLOPE with $p = 26,315$ groups, where each of the groups contains two explanatory variables, for the additive and the dominance effect of the same SNP,

A.2 SLOPE$_X$, where the regular SLOPE is used to search through the reduced design matrix X (as in Bogdan et al., 2015; Brzyski et al., 2017),

A.3 SLOPE$_{XZ}$, where the regular SLOPE is used to search through the full design matrix $[X\ Z]$.

In all versions of SLOPE we used the iterative procedure for estimation of σ and the sequence λ heuristically adjusted to the case of the Gaussian design matrix, as implemented in the CRAN packages `SLOPE` and `grpSLOPE`. All scripts used in simulations as well as in real data analysis are available at `https://github.com/dbrzyski/gSLOPE`.

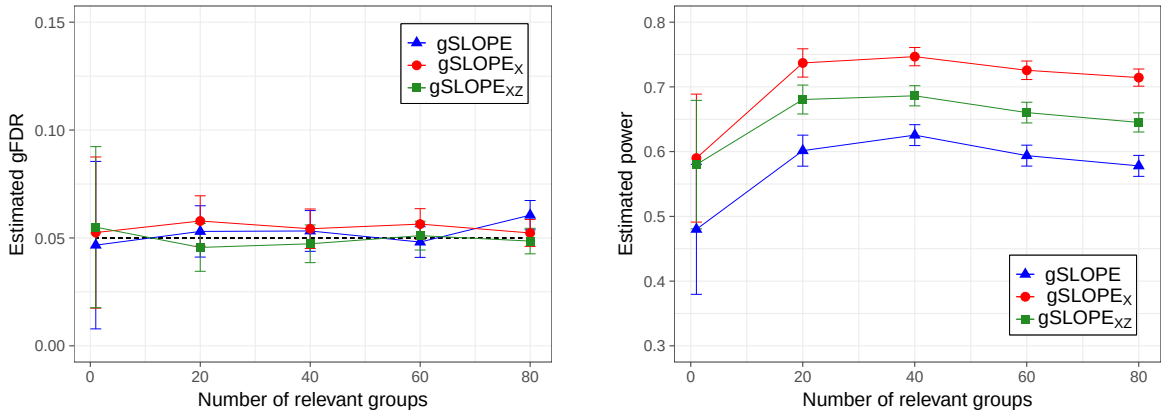
Figure 5 summarizes this simulation study. Here FDR and power are calculated at the SNP level. Specifically, in case of SLOPE$_{XZ}$ the SNP is counted as a one discovery if the corresponding additive or the dominance dummy variable is selected.

As shown in Figure 5, for both of the simulated scenarios all versions of SLOPE control gFDR for all considered values of k. When the data are simulated according to the additive model the highest power is offered by SLOPE$_X$, with the power of gSLOPE being smaller by approximately 13% over the whole range of k. However, in the presence of large dominance effects the situation is reversed and gSLOPE offers the highest power, which systematically exceeds the power of SLOPE$_X$ by the symmetric amount of 13%. In our simulations SLOPE$_{XZ}$ has intermediate performance and does not substantially improve the power of SLOPE$_X$ in the presence of dominance effects.

Thus our simulations suggest that gSLOPE provides an information complementary to SLOPE$_X$ and our recommendation is to use both these methods when performing GWAS. SNPs detected by gSLOPE and not detected by SLOPE$_X$ almost certainly exhibit strong dominance effects and might represent rare recessive variants, as suggested by the real data analysis reported in the following section.

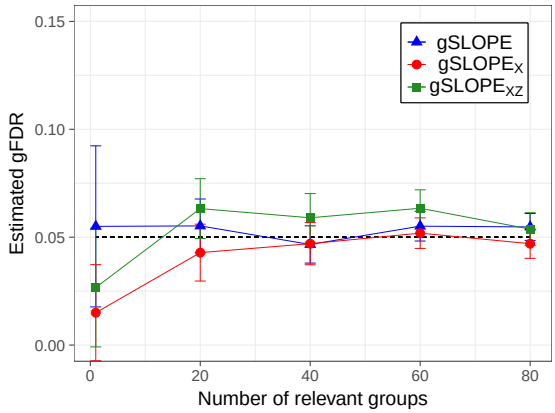
2.9 gSLOPE under GWAS application: real phenotype data

Finally, we have applied group SLOPE to identify SNPs associated with four lipid phenotypes available in NFBC1966 dataset. This data set contains many characteristics of

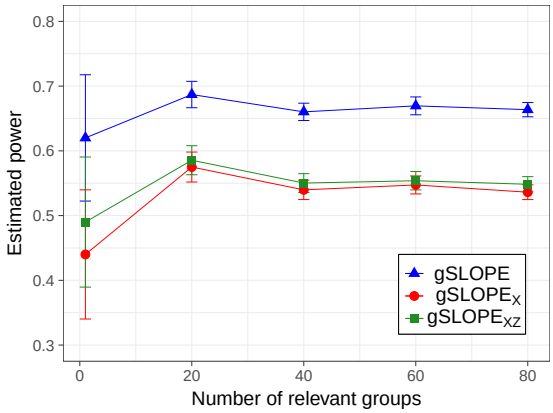


(a) FDR, additive

(b) Power, additive



(c) FDR, dominance



(d) Power, dominance

Figure 5: Simulations using real SNP genotypes: $n = 5,402$, $p = 26,315$. Power and gFDR are estimated based on 100 iterations of each simulation scenario. Upper panel illustrates the situation where all causal SNPs have only additive effects, while in lower panel each causal SNP has also some dominance effect.

individuals from the Northern Finland Birth Cohort 1966 (NFBC1966) ((Rantakallio, 1969; Jarvelin et al., 2004)), a sample that enrolled almost all individuals born in 1966 in the two northernmost Finnish provinces. The most advantageous feature of this study is that "participants derive from a genetic isolate that is relatively homogeneous in genetic background and environmental exposures and that has more extensive linkage disequilibrium (i.e. neighboring markers are more strongly correlated) than in most other populations"(see Sabatti et al., 2009). The second of these features allows to cap-

ture the associations resulting from mutations which are not genotyped (i.e. they are represented in the design matrix only through their neighbors). In (Sabatti et al., 2009) this data set was used to look for associations for "nine quantitative traits that are heritable risk factors for cardiovascular disease (CVD) or type 2 diabetes (T2D): body mass index (BMI), fasting serum concentrations of lipids (triglycerides (TG), high-density lipoproteins (HDL) and low-density lipoproteins (LDL)), indicators of glucose homeostasis (glucose (GLU), and insulin (INS)) and inflammation (CRP), and systolic (SBP) and diastolic (DBP) blood pressure. Extreme values of these traits, in combination, identify a metabolic syndrome, hypothesized to increase risks for both CVD and T2D". In (Brzyski et al., 2017) four lipid phenotypes: HDL, LDL, TG, and total cholesterol (CHOL), were reanalyzed using the `geneSLOPE` method based on SLOPE$_X$. The results were compared to those obtained with the up-to-date EMMAX procedure (Kang et al., 2010), which controls for the polygenic background by using the mixed model approach. The study reported in (Brzyski et al., 2017) shows that in this example `geneSLOPE` usually points at the same genomic regions as EMMAX, but allows to obtain a better resolution of gene location. Here we analyze the same four traits (HDL, LDL, TG, and CHOL) with the `geneSLOPE` based on SLOPE$_{XZ}$ and group SLOPE and compare the results with those obtained with SLOPE$_X$ and reported in (Brzyski et al., 2017).

We started with $5,402$ individuals and $334,103$ SNPs, obtained after first step of screening procedure described in details in Appendix I. Since this pre-processing selects most promising SNPs by performing multiple testing on the full set of $p = 334,103$ SNPs, the sequence of the tuning parameters for SLOPE needs to be adjusted to this value of p rather than to the number of selected representatives (see Brzyski et al., 2017). The algorithm for GWAS analysis with SLOPE (the entire procedure is called geneSLOPE) is implemented in `R` package `geneSLOPE` and its details are explained

in (Brzyski et al., 2017). According to an extensive simulation study and real data analysis reported in (Brzyski et al., 2017), `geneSLOPE` allows to control FDR for the analysis with full size GWAS data.

In our data analysis we used three methods: geneSLOPE for additive effects (as in Brzyski et al., 2017), geneSLOPE$_{XZ}$, with the design matrix extended by inclusion of dominance dummy variables, and gene group SLOPE (geneGSLOPE). In geneSLOPE$_{XZ}$ and geneGSLOPE representative SNPs were selected based on one-way ANOVA tests. For all these procedures the pre-processing was based on p-value threshold $p < 0.05$ and the correlation cutoff $\rho < 0.3$, which allowed to reduce the data set to roughly 8500 of interesting representative SNPs (this number depends on the phenotype). For the convenience of the reader, the Procedure 5 for the full geneGSLOPE analysis is provided below.

Procedure 5 geneGSLOPE procedure

Input: $r \in (0, 1)$, $\pi \in (0, 1]$

Screen SNPs:
(1) For each SNP calculate independently the p-value for the ANOVA test with the null hypothesis, $H_0 : \mu_{aa} = \mu_{aA} = \mu_{AA}$.
(2) Define the set $\mathcal{B}$ of indices corresponding to SNPs whose p-values are smaller than π.

Cluster SNPs:
(3) Select the SNP j in $\mathcal{B}$ with the smallest p-value and find all SNPs whose Pearson correlation with this selected SNP is larger than or equal to r.
(4) Define this group as a cluster and SNP j as the representative of the cluster. Include SNP j in $\mathcal{S}$, and remove the entire cluster from $\mathcal{B}$.
(5) Repeat steps (3)-(4) until $\mathcal{B}$ is empty. Denote by m number of all clumps (this is also the number of elements in $\mathcal{S}$).

Selection:
(6) Apply the iterative gSLOPE method (i.e. gSLOPE with σ estimation and correction for independent regressors) on $X_{\mathcal{S}}$, being matrix X restricted to columns corresponding to the set $\mathcal{S}$ of selected SNPs. Here, the tuning parameters, vector λ, is defined as in Procedure 3, with p being the number of all initial SNPs, and then this vector is restricted only to first m coefficients.
(7) Representatives which were selected indicate the selection of entire clumps.

Results in the context of number of discoveries given by geneSLOPE, geneSLOPE$_{XZ}$ and geneGSLOPE are summarized in Table 2, where we can observe that both geneSLOPE and geneSLOPE$_{XZ}$, gave identical results for LDL, CHOL and TG. Compared to these methods geneGSLOPE did not reveal any new response-related SNPs for LDL

and CHOL. Actually, for these two traits geneGSLOPE missed some SNPs detected by the other two methods. A different situation takes place for TG, where geneGS-

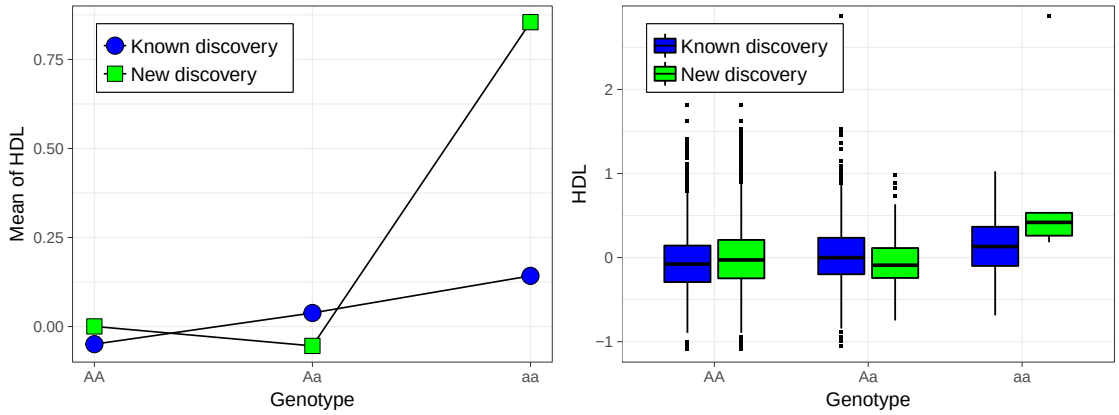
	HDL	LDL	TG	CHOL
geneSLOPE	7	6	2	5
geneSLOPE$_{XZ}$	8	6	2	5
geneGSLOPE	8	4	8	4
New discoveries: geneSLOPE$_{XZ}$	1	0	0	0
New discoveries: geneGSLOPE	2	0	6	0

Table 2: Number of discoveries in real data analysis

LOPE identifies 6 additional SNPs as compared to the other two methods. All these detections have a similar structure, showing a significant recessive effect of the minor allele. In all these cases the minor allele frequency was smaller than 0.1. The detection of such "rare" recessive effects by the simple linear regression model is rather difficult, since the regression line adjusts mainly to the two prevalent genotype groups and is almost flat (Lettre et al., 2007).

In case of HDL all three versions of SLOPE gave different results. geneSLOPE$_{XZ}$ identifies one new SNP as compared to geneSLOPE, while geneGSLOPE identifies one more SNP and misses one of the discoveries obtained by other two methods. In Figure 6 we compare two exemplary discoveries: one detected at the same time by geneSLOPE and geneGSLOPE (known discovery) and one detected only by geneGSLOPE (new discovery). This example clearly shows the additive effect of the previously detected SNP and the recessive character of the second SNP. In case of new discovery there are only 5 individuals in the last genotype group, which makes the change in the mean not detectable by simple linear regression.

The results of real data analysis agree with results of simulations. They show that geneGSLOPE has a lower power than geneSLOPE for detection of additive effects but can be very helpful in detecting rare recessive variants. Thus these two methods are



(a) Mean of HDL for different genotypes (b) Boxplots of HDL for different genotypes

Figure 6: Comparison of a discovery detected by both geneSLOPE and geneGSLOPE (known discovery), and a discovery detected only by geneGSLOPE (new discovery). The mean values of HDL for different genotypes are shown in (a) and the corresponding boxplots are presented presented in (b).

complementary to each other and should be used together to enhance the power of detection of influential genes.

3 Discussion

Group SLOPE is a new convex optimization procedure for selection of important groups of explanatory variables, which can be considered as a generalization of group LASSO and of SLOPE. In this article we provide an algorithm for solving group SLOPE and discuss the choice of the sequence of regularizing parameters. Our major focus is the control of group FDR, which can be obtained when variables in different groups are orthogonal to each other or they are stochastically independent and the signal is sufficiently sparse. After some pre-processing of the data such situations occur frequently in the context of genetic studies, which in this paper serve as a major example of applications.

The major purpose of controlling FDR rather than absolutely eliminating false discoveries is the wish to increase the power of detection of signals which are comparable

to the noise level. As shown by a variety of theoretical and empirical results, this allows SLOPE to obtain an optimal balance between the number of false and true discoveries and leads to very good estimation and predictive properties (see e.g. Bogdan et al., 2013, 2015; Su and Candès, 2016). Our Theorem 2.6 illustrates that these good estimation properties are inherited by group SLOPE.

We provide the regularizing sequence $\lambda^{\max}$, which provably controls gFDR in case when variables in different groups are orthogonal. Additionally, we propose its relaxation λ^{mean}, which according to our extensive simulations controls "average" gFDR, where the average is with respect to all possible signal placements. This sequence can be easily modified taking into account the prior distribution on the signal placement. Such "Bayesian" version of gSLOPE and the proof of control of the respective average gFDR remains an interesting topic for a further research.

Another important topic for a further research is the formal proof of gFDR control when variables in different groups are independent and setting precise limits on the sparsity levels under which it can be done. Asymptotic formulas, which allow for accurate prediction of FDR for LASSO under Gaussian design are provided in (Su et al., 2015). We expect that similar results can be obtained for SLOPE and gSLOPE and generalized to the case of random matrices, where variables are independent and come from sub-Gaussian distributions. We consider this as an interesting topic for a further research.

While we concentrated on control of FDR in case when groups of variables are roughly orthogonal to each other, it is worth mentioning that original SLOPE has very interesting properties also in case when regressors are strongly correlated. As shown e.g. in (Figueiredo and Nowak, 2016), the Sorted L-One norm has a tendency to average estimated regression coefficients over groups of strongly correlated predictors, which enhances the predictive properties. This also allows not to lose important pre-

dictors due to their correlation with other features. Moreover, minimax estimation and prediction properties of SLOPE under correlated designs have been recently proved in (Bellec et al., 2016b) and (Bellec et al., 2016a). We expect similar properties to hold for gSLOPE, which would pave the way for the applications in a variety of applications, where the groups of predictors are not necessarily independent.

Our proposed construction of the group SLOPE allows for the estimation of the group effects but does not allow to estimate the regressor coefficients by individual explanatory variables. We believe that the estimation of the individual effects would require a modification of the penalty term, so that a penalty would be imposed not only on entire groups, but also on individual coefficients. Such an idea was used in (Simon et al., 2013) in the context of sparse-group LASSO, where an additional l_1 penalty on individual coefficients was used. The modification of gSLOPE in this direction would be an interesting contribution, since it could be applied for a bi-level selection the selection of groups and particular variables within the selected groups at the same time. However, achieving gFDR control with such a modified penalty currently seems to be a challenging task.

We proposed a specific application of gSLOPE for Genome Wide Association Studies, where groups contain different effects of the same SNP. It is also worth mentioning that gSLOPE can be used to group SNPs based on biological function, physical location etc. We also expect this method to be advantageous in the context of identification of groups of rare genetic variants, where considering their joint effect on phenotype should substantially increase the power of detection. Going beyond genetic analysis group SLOPE could be used also for example in neuroimaging studies, where one can group voxel-wise brain activity measures, such as the ones derived from functional magnetic resonance imaging (fMRI), using the region of interest (ROI) definitions given by available anatomical atlases. Apart from these bioinformatics and medicine applica-

tions, one could also consider application of a group SLOPE for a variety of compressed sensing tasks. Here the most basic application of block/group sparsity is the extension to complex numbers in which the real and imaginary parts are split (and all represented by their coefficient), with each pair forming a group (see e.g. van den Berg and Friedlander, 2008; Maleki et al., 2013). As discussed e.g. in (Elhamifar and Vidal, 2012) block sparsity arises also in a variety of other applications such as reconstructing multi-band signals, face/digit/speech recognition or clustering of data on multiple subspaces etc (see Elhamifar and Vidal, 2012, for the respective references). Another interesting application discussed in (van den Berg and Friedlander, 2011) is the identification of the temporal signals arriving from different, unknown, but stationary directions. Here the sparsity is the direction of arrival, whereas the L_2-norm is over the time series corresponding to this direction. These few possible applications represent only a small part of the real life group sparsity scenarios and exciting potential applications for the group SLOPE and we look forward investigating practical properties of group SLOPE in these real life problems.

Acknowledgement

We would like to thank Ewout van den Berg, Emmanuel J. Candès and Jan Mielniczuk for helpful remarks and suggestions. D. B. would like to thank Professor Jerzy Ombach for significant help with the process of obtaining access to the data. D. B. and M. B. are supported by European Union's 7th Framework Programme for research, technological development and demonstration under Grant Agreement no 602552 and by the Polish Ministry of Science and Higher Education according to agreement 2932/7.PR/2013/2. Additionally D.B. acknowledges the support from NIMH grant R01MH108467.

References

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 1:183–202.

Bellec, P., Lecué, G., and Tysbakov, A. (2016a). Bounds on the prediction error of penalized least squares estimators with convex penalty. *arXiv:1609.06675*.

Bellec, P., Lecué, G., and Tysbakov, A. (2016b). Slope meets lasso: improved oracle bounds and optimality. *arXiv:1605.08651*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

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

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope – adaptive variable selection via convex optimization. *Annals of Applied Statistics*, 9(3):1103–1140.

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

Brzyski, D., Peterson, C., Sobczyk, P., Candès, E., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of gwas false discoveries. *Genetics*, 205:61–75.

Brzyski, D., Su, W., and Bogdan, M. (2015). Group slope - adaptive selection of groups of predictors. *arXiv:1310.1969*.

Elhamifar, E. and Vidal, R. (2012). Block-sparse recovery via convex optimization. *IEEE Transactions on Signal Processing*, 60(8):4094–4107.

Figueiredo, M. A. T. and Nowak, R. D. (2016). Ordered weighted l_1 regularized regression with strongly correlated covariates: Theoretical aspects. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, JMLR:W&CP*, 51:930–938.

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

Gossmann, A., Cao, S., and Wang, Y.-P. (2015). Identification of significant genetic variants via SLOPE, and its extension to group SLOPE. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*.

Jarvelin, M., Sovio, U., King, V., Lauren, L., Xu, B., McCarthy, M., Hartikainen, A., Laitinen, J., Zitting, P., Rantakallio, P., and Elliott, P. (2004). Early life factors and blood pressure at age 31 years in the 1966 northern finland birth cohort. *Hypertension*, 44:838–846.

Kang, H., Sul, J., Service, S., Zaitlen, N., Kong, S., Freimer, N., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–355.

Lettre, G., Lange, C., and Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31(4):358–362.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. (2013). Asymptotic analysis of complex lasso via complex approximate message passing (camp). *IEEE Transactions on Information Theory*, 59(7):4290–4308.

Neuvial, P. and Roquain, E. (2012). On false discovery rate thresholding for classification under sparsity. *Annals of Statistics*, 40:2572–2600.

Purcell, S. (2009). package plink.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81.

Rantakallio, P. (1969). Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr. Scand. Suppl.*, 193:43.

Sabatti, C., Service, S. K., Hartikainen, A., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N. A., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot,

P., McCarthy, M. I., Daly, M. J., Jrvelin, M., Freimer, N. B., and Peltonen, L. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, 41(1):35–46.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Simon, N. and Tibshirani, R. (2013). Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983–1001.

Su, W., Bogdan, M., and Candès, E. (2015). False discoveries occur early on the lasso path. *arXiv:1511.01957, to appear in Ann. Statist.*

Su, W. and Candès, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Annals of Statistics*, 40:1038–1068.

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

Tseng, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington.

van den Berg, E. and Friedlander, M. P. (2008). Probing the pareto frontier for basis pursuit solutions. *SIAM J. on Scientific Computing*, 31(2):890–912.

van den Berg, E. and Friedlander, M. P. (2011). Sparse optimization with least-squares constraints. *SIAM J. on Optimization*, 21(4):1201–1229.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67.