



SELECTING RELEVANT GROUPS OF
EXPLANATORY VARIABLES VIA CONVEX
OPTIMIZATION METHODS WITH THE FALSE
DISCOVERY RATE CONTROL

Damian Brzyski

A thesis submitted for the degree of Doctor of
Philosophy

Supervisor of the thesis is prof. Małgorzata Bogdan
from University of Wrocław

October 2015

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Aims and Key Results	4
1.3	Outline	5
2	Background	5
2.1	Model selection problem	5
2.2	Using SLOPE to control FDR	6
2.3	Group selection in linear regression	8
3	Notations, mathematical tools and techniques	10
3.1	Permutations and rearrangement inequality	11
3.2	Optimization problems and duality	11
3.3	Convexity	13
3.4	Subgradients and KKT conditions	14
3.5	Proximal gradient method	15
3.6	Matrix norms and random matrices	16
4	J_λ norm and SLOPE properties	17
4.1	J_λ norm properties	17
4.2	SLOPE with diagonal experiment matrix	18
5	group SLOPE	21
5.1	Formulation of the optimization problem	21
5.2	Standardization of design matrix	23
5.3	Orthogonal situation	24
5.4	Strength of signals	28
5.5	The impact of chosen weights	29
5.6	Near-orthogonal situation	31
5.7	The estimation of the variance of stochastic error, σ^2	36
6	Numerical algorithm	37
7	GWAS simulations	37
8	Group Ordered Dantzig Selector	40
A	gFDR control with gSLOPE under orthogonal case	51
A.1	The proof of Lemma 1	51
A.2	The proof of Lemma 3	52
B	The expected values of random matrices	53
B.1	The proof of Lemma 4	53
B.2	The proof of Lemma 5	53
C	Stopping criteria for numerical algorithm	54

1 Introduction

1.1 Motivation

Assume that the n by p dimensional experiment (design) matrix, X , is given and consider the classical multiple regression model

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

where $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Moreover, assume that the vector of values of the response variable, y , is provided, while β is unknown vector of regression coefficients. The purpose of the statistical analysis is to recover the support of β , which identifies the set of important regressors. In many applications number of observations, n , is much smaller than p . In such situations the assumption that number of nonzero coefficients in β is relatively small comparing to p , is crucial.

The above problem is known as the model selection problem and there exist many different methods applicable to this issue, which has drawn increasing attention recently. The usual approach is to define estimate as the solution of minimization problem with the objective $\|y - Xb\|_2^2 + \lambda f(b)$, where in typical applications $f(b)$ is the number of nonzero elements of b or some norm inducing sparse solutions and λ is vector of tuning (smoothing) parameters. Such an estimation could be treated as a compromise between fit and complexity.

Such methods of model selection mostly need some particular selection for tuning parameters and this choice is found to be very challenging. The general rule is that when one reduces the tuning parameter, then the method can identify more elements from true support (true discoveries) but at the same time it produces more falsely identified variables (false discoveries). The natural approach is to decide on the some maximal fraction of expected value of false discoveries among all discoveries and choose smoothing parameters in such way that this value is not exceeded. We refer to the ability of selecting such parameters as achieving the False Discovery Rate control (FDR control).

In this thesis we will focus on methods allowing to achieve FDR control in situation when considered data have some special structure, i.e. one can divide explanatory variables into small groups such that variables coming from different groups are weakly correlated. The motivation comes directly from the genome-wide association studies (GWAS), which rely on the identification of single nucleotide polymorphisms (SNPs) associated with some quantitative trait, e.g. height, gene expression or blood pressure. The standard approach is to store coded genotypes (i.e. different values for different variants of SNPs) in matrix $X \in M(n, p)$, where p is the number of all considered SNPs and n represents the number of all individuals involved in the study. GWAS are most commonly presented as a problem of explanatory variables selection in the linear regression model. In this context the major issue which needs to be addressed is the statistical correlation between genotypes of neighboring SNPs. As illustrated in Figure 1 the matrix of SNP correlations has a block diagonal structure, with blocks corresponding to some conserved regions of the genome.

Such situation (the occurrence of high correlations) can lead to worse performance of statistical tools compared to an uncorrelated case [1]. If dependencies between different variables are strong, the information about true support, which is contained in collected observations, could be easily disturbed, even with high signal-to-noise ratio. This effect may results in producing a relatively large amount of false discoveries. In particular, we will show that recently proposed method, *Sorted L-One Penalized Estimation* (SLOPE) [2], designed to select relevant regressors with control over FDR, fails when is applied directly to raw GWAS data.

In many situations the distinction between strongly correlated regressors becomes statistically pointless. It seems much more reasonable to give up the idea of selecting one true regressor and instead concentrate on selecting the entire group of closely correlated SNPs which are associated with the trait and point to the region hosting a causal mutation. The idea directly leads to such concepts and terms as response related groups selection, truly/false discovered groups or FDR control at group level, which we are investigating in this thesis.

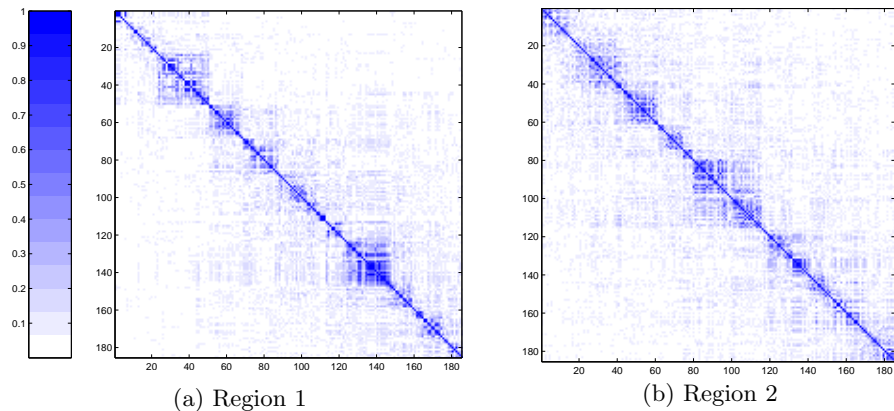


Figure 1: Heat maps of correlation matrices (absolute values) for consecutive columns from design matrix corresponding to two separated regions in chromosome 1

1.2 Aims and Key Results

The overall aim of this thesis is to derive new methods for explanatory variables selection in linear regression having property of FDR control (at the groups context) under assumptions concerning the structure of considered data (in terms of correlations between explanatory variables). Specifically we aim to:

- introduce the concept of groups selection with FDR control at groups level;
- propose extension of SLOPE, capable to achieve such control under certain assumptions concerning the structure of data;
- derive efficient numerical solver, capable to obtain estimates even for large dimensions of design matrices;
- perform experiments with generated and real genetic data;
- consider the respective generalization of the Dantzig Selector method.

Key results of this thesis are listed below.

- We introduce group SLOPE (gSLOPE) method applicable to groups selection problem, defining estimate as solution to a specific convex optimization problem;
- we derive a fast algorithm, based on proximal gradient method, for finding numeric solution of this problem;
- we give precise procedure for choosing starting parameters such that FDR control at groups level is ensured under orthogonal design for arbitrary sequence of groups sizes;
- we give heuristic procedure of choosing starting parameters in near-orthogonal situation and illustrate its good performance with numerically generated data;
- we incorporate the estimation of stochastic error variance to our method;
- we use final procedure in experiments with real genetic data showing potential of group SLOPE in genetic application;
- additionally, we introduce the alternative method for groups selection, the group Ordered Dantzig Selector (gODS), which defines estimate as solutions to specific constrained convex optimization problem;

- we investigate the single-variable-in-group scenario (which reduces groups selection problem to classical model selection problem) and show that the solutions to gODS and gSLOPE are identical under orthogonal situation and strictly monotonic sequence of tuning parameters. As a result we give procedure for choosing these parameters such that FDR control for gODS is ensured.

1.3 Outline

The thesis structure is as follows.

In Chapter 2 we provide some background of model selection problem and we discuss the issue of FDR control. The special attention is given to the most recently developed method, SLOPE, which provides the basis for results contained in this thesis - in particular we present the main theoretical result from [2] concerning SLOPE properties in the context of FDR control under orthogonal situation. Moreover, in this chapter we discuss the issue of selecting entire groups of explanatory variables in linear regression models. In Chapter 3 we list the most often used notations and recall necessary definitions and theorems, which will be used in further part of this thesis. The next Chapter is devoted to sorted ℓ_1 norm and SLOPE method. We present interesting properties of this norm, concerning SLOPE with diagonal design matrix and we generalize lemmas proved in [2]. Tools introduced in this Chapter are used in Chapter 5, where we present the main contribution of the thesis - the group SLOPE method. In this Chapter we formulate and prove the main theoretical result for this method - the theorem about choosing smoothing parameters under orthogonal situation. Furthermore, we derive the heuristic procedure for using gSLOPE in near-orthogonal case and present results of Monte Carlo simulations. In Chapter 6, we describe the use of the proximal gradient method to obtain the numerical algorithm for finding group SLOPE estimates. Experiments with real genetic data are performed in Chapter 7, where we use gSLOPE to select entire groups of correlated predictors, aiming to control the fraction of false discoveries. In Chapter 8 we propose the alternative methods for variables and groups selecting - the group Ordered Dantzig Selector and the Ordered Dantzig Selector. We prove theorem about equivalence between SLOPE and ODS under orthogonal situation. Some proofs and technical aspects connected with the numerical algorithms are placed in the Appendix.

2 Background

2.1 Model selection problem

Model selection problem in linear regression is the task of selecting a subset of explanatory variables such as this choice is the best in terms of given criteria (therefore it is often referred to as the problem of subset selection). The problem has been widely studied and various approaches can be found in the literature. A common procedure is to use the minimization problem with an objective function defined as the weighted sum of two components: the term responsible for the goodness of fit and the second term, which is a penalty for the number of variables included to the model. Such approaches, known as penalized likelihood criteria, are represented by the Akaike Information Criterion (AIC) [3] and the Bayesian Information Criterion (BIC) [4].

The drawback of these methods is that minimizing over all possible subsets gives 2^p possible combinations. Even after restricting only to a subset of models having maximally $k < p$ variables, number of all possible choices is still astronomic in practical problems. This makes it not feasible to compute model selection criteria for all of them. More recently, an alternative class of penalization methods, represented by the Least Absolute Shrinkage and Selection Operator (LASSO) [5], has become popular. The idea is to consider some sparsity inducing norm (instead of a penalty based on cardinality of model) and defining the estimate as a support of solution to a convex optimization problem, i.e. the set of nonzero coefficients of solution corresponds to variables identified as relevant.

LASSO uses sparsity enforcing properties of ℓ_1 norm for this purpose, which yields problem

$$\beta^L := \arg \min_b \left\{ \frac{1}{2} \|y - Xb\|^2 + \lambda_L \sigma \|b\|_1 \right\}, \quad (2.1)$$

where λ_L is a tuning parameter defining the trade-off between the fitting of model and the sparsity of solution. It is known that in wide range of situations, the LASSO can select the optimal subset of variables and that it achieves a squared error within a logarithmic factor of the ideal mean squared error (i.e. error that one would achieve with the knowledge about true support) [6]. These, among others, are the reasons why there has been much work in recent years generalizing the LASSO to a variety of problems and applying this method in many fields, including statistics, engineering, mathematics and computer science. More or less connected with LASSO are such approaches as *the Elastic Net* [7], *The Adaptive Lasso* [8], *the Least Angle Regression* [9], *Relaxed Lasso* [10] or SLOPE (discussed in next subsection), to name just a few.

The method which is also related with LASSO is *the Dantzig Selector* (DS) [11]. DS minimizes the ℓ_1 norm subject to an ℓ_∞ constraint on the correlation of the residuals with the predictors, i.e. estimate is defined as the solution to constrained optimization problem

$$\beta^{DS} := \arg \min_b \|b\|_1 \quad \text{subject to} \quad \|X^T(y - Xb)\|_\infty \leq \sigma \lambda_{DS}, \quad (2.2)$$

where $\lambda_{DS} > 0$ is a sparsity dictating parameter. The method was introduced by Candès and Tao in 2007 and it shed a new light on high-dimensional model selection. The important advantage is the fact that DS can be simply recast as a linear program and solved usually faster than the existing methods by using one of many available algorithms for such problems. Moreover, authors showed that the DS is very accurate - under certain assumptions about the matrix X , with large probability DS mimics the risk of the oracle estimator up to a logarithmic factor of p .

2.2 Using SLOPE to control FDR

The concept of using the false discovery rate, as a appropriate criterion in identifying the important effects, was proposed by Benjamini and Hochberg [12] in 1995. The idea was introduced in the context of multiple testing problem, i.e. situation when statistical inferences are considered simultaneously. Such a problem occurs for example when one identifies significant regressors by performing simple regression tests, separately at every SNP. In case of such a multiple testing problem the decision is based on p -dimensional vector of p-values. Here we will discuss mainly the Benjamini-Hochberg procedure (BH, [12]), which has been proved to control the expected proportion of false discoveries among all discoveries below any predefined level $\alpha \in [0, 1]$, if test statistics are independent. The procedure (the step-up version) works as follows

Procedure 1 Benjamini-Hochberg procedure

1. order the p-values in non-decreasing sequence $P_1 \leq \dots \leq P_p$;
 2. for a given α , find the largest k such that $P_k \leq \frac{k}{p} \alpha$;
 3. declare discoveries for all tests with p-values smaller or equal than P_k .
-

There exists strong connection between multiple testing and model selection [13]. This is no straightforward, however, how to apply BH procedure for a multivariate linear model. To fill this gap, the extension of LASSO, SLOPE, was recently introduced. Instead of the ℓ_1 norm (as in LASSO case), the method uses FDR control properties of J_λ norm defined below.

For a sequence lambda satisfying $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ and $b \in \mathbb{R}^p$, the *sorted ℓ_1 norm* is given as

$$J_\lambda(b) := \sum_{i=1}^p \lambda_i |b|_{(i)}, \quad (2.3)$$

where $|b|_1 \geq \dots \geq |b|_p$ are sorted absolute values. SLOPE is the solution to convex optimization problem

$$\beta^s := \arg \min_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \sigma J_\lambda(b) \right\}, \quad (2.4)$$

which clearly reduces to LASSO problem, (2.1), for $\lambda_1 = \dots = \lambda_p =: \lambda_L$.

Similarly as in classical model selection, the support of solution defines the subset of variables estimated as relevant. It was shown in [2] that SLOPE is strongly connected with BH procedure under orthogonal case, i.e. when $X^\top X = \mathbf{I}_p$. Below we present the main theoretical result from this article, which is crucial for further contents of this thesis.

Theorem 1 (Theorem 1.1 of [2]). *Let q be any number from $(0, 1)$ and $\beta \in \mathbb{R}^p$ be unknown vector. Denote the number of zero coefficients in β by p_0 and assume that:*

- $X^\top X = \mathbf{I}_p$,
- $Y = X\beta + z$, with $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$,
- y is a given realization of random vector Y ,
- $\lambda^{\text{BH}} = (\lambda_1^{\text{BH}}, \dots, \lambda_p^{\text{BH}})^\top$, for $\lambda_i^{\text{BH}} := \Phi^{-1}\left(1 - \frac{q_i}{2p}\right)$, where Φ is a cumulative function of the standard normal distribution
- β^s is solution to

$$\arg \min_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \sigma J_{\lambda^{\text{BH}}}(b) \right\},$$

- V and R are defined as $V := \left| \{i : \beta_i = 0, \beta_i^s \neq 0\} \right|$, $R := \left| \{i : \beta_i^s \neq 0\} \right|$.

Then, it holds

$$FDR := \mathbb{E} \left[\frac{V}{\max\{R, 1\}} \right] \leq q \cdot \frac{p_0}{p}.$$

Moreover, in [14] it is proved that SLOPE is also very accurate in terms of prediction. Let denote by $\|\beta\|_0$ the number of nonzero entries of β . With the sequence λ^{BH} , SLOPE adapts to unknown sparsity and is asymptotically minimax under orthogonal (and under random Gaussian design, described later in (2.6)), meaning that

$$\sup_{\|\beta\|_0 \leq k} \mathbb{E} \|\beta^s - \beta\|_2^2 \sim \inf_{\hat{\beta}} \sup_{\|\beta\|_0 \leq k} \mathbb{E} \|\hat{\beta} - \beta\|_2^2, \quad (2.5)$$

where $a \sim b$ denotes the situation when $a/b \rightarrow 1$ while $k/p \rightarrow 0$, and the infimum is taken over all measurable estimators.

Clearly, it is highly unlikely that the idealistic assumption about orthogonality of columns in the design matrix could be considered as valid, when one works with real-world data. This is the reason why choosing proper parameters for SLOPE in general situation is very challenging. However, there exist applications in which assumption about relatively weak correlations between different columns is reliable. In such situation it is often sensible to assume that the design matrix (after centering and normalizing columns) could be modeled as a realization of random matrix in which entries are independent zero-mean random variables. A specific and important example is a Gaussian design where

$$X_{ij} \sim \mathcal{N}(0, 1/n), \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, p\}. \quad (2.6)$$

Clearly, then it also occurs $\mathbb{E}(X_i^\top X_j) = 0$ for $i \neq j$ and $\mathbb{E}(X_i^\top X_i) = 1$.

In this thesis, we will refer to (2.6) as the near-orthogonal situation. Heuristic about selecting starting parameter under (2.6) was derived in [2]. We will now recall this procedure, since it is connected with the reasoning introduced in this thesis.

Procedure 2 Selecting lambdas for SLOPE in near-orthogonal situation [2]

Input: $q \in (0, 1)$, p , $n \in \mathbb{N}$

1. set $\lambda^{\text{BH}} = (\lambda_1^{\text{BH}}, \dots, \lambda_p^{\text{BH}})^\top$, for $\lambda^{\text{BH}}(i) := \Phi^{-1}\left(1 - \frac{qi}{2p}\right)$;
2. define

$$\lambda_i^{\text{G}} := \begin{cases} \lambda_1^{\text{BH}} & , i = 1 \\ \lambda_i^{\text{BH}} \sqrt{1 + \sum_{j < i} \frac{(\lambda_j^{\text{G}})^2}{n-i}} & , i > 1 \end{cases} ;$$

3. find the largest index, k^* , such that $\lambda_1^{\text{G}} \geq \dots \geq \lambda_{k^*}^{\text{G}}$;
4. put

$$\lambda_i := \begin{cases} \lambda_i^{\text{G}} & , i \leq k^* \\ \lambda_{k^*}^{\text{G}} & , i > k^* \end{cases} .$$

Computer simulations and real data analysis reported in [2] show that Procedure 2 generates the sequence of lambdas which allows for FDR control for a variety of common random design matrices. Good performance of this method was also observed in situation when non-Gaussian, zero-mean distribution is used to generate the entries of design matrix.

2.3 Group selection in linear regression

There are many statistical problems and applications in which the idea of grouping variables and selecting entire groups, rather than individual predictors, naturally appears. One of the most well-known examples is multifactor analysis-of-variance problem where the goal is to find response-related explanatory factors, while each of them represents a group of variables.

In the problem of group selection one aims at identifying groups of variables which are response-related. Assume that the division into groups is defined by the family $I = \{I_1, \dots, I_m\}$ consisting of m subsets forming the partition of the set $\{1, \dots, p\}$ and rewrite (1.1) as the linear regression model with m groups of form

$$y = \sum_{i=1}^m X_{I_i} \beta_{I_i} + z, \tag{2.7}$$

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 . As before we assume that all columns have unit ℓ_2 norm.

In the classical multiple regression model (1.1), the effect size of i th variable could be simply defined as the magnitude of β_i . The association is clear: the support of β is exactly the same as indices corresponding to relevant variables (defined as variables with nonzero effect sizes). In the context of group selection it seems natural to measure the group effects by the values $\|\beta_{I_1}\|_2, \dots, \|\beta_{I_m}\|_2$ and estimate support of vector $\llbracket \beta \rrbracket_I := (\|\beta_{I_1}\|_2, \dots, \|\beta_{I_m}\|_2)^\top$. Such a definition, however, does not take into account the dependency structure inside groups. In particular, when columns in group i are linearly dependent, it may happen that $X_{I_i} \beta_{I_i} = 0$ while not all coefficients of β_{I_i} are zeros. Defining truly relevant groups by the occurrence of nonzero coefficients in β_{I_i} would result in labeling such group as truly relevant which makes no sense.

In this thesis we will use different approach: we will treat the value $\|X_{I_i} \beta_{I_i}\|_2$ as a measure of impact of i th group on the response (we will also use terms *the effect size* or *the strength* of i th group) and we will say that group i is truly relevant if and only if $\|X_{I_i} \beta_{I_i}\|_2 > 0$. Consequently, we are focusing on estimating the support of $\llbracket \beta \rrbracket_{X,I} := (\|X_{I_1} \beta_{I_1}\|_2, \dots, \|X_{I_m} \beta_{I_m}\|_2)^\top$. There are many reasons why one should consider such a definition of group effect strength, rather than ℓ_2 norm imposed directly on β_{I_i} . Consider the simple, intuitive example when columns included in different groups are orthogonal to each other and $y = \sum_{i=1}^m X_{I_i} \beta_{I_i}$ (the noiseless case). Moreover, assume that columns were centered, meaning that the sum of coefficients in each X_i is equal to

zero. The impact of the i th group could be viewed through the dependence between y and the i th component of sum, $\alpha_i := X_{I_i}\beta_{I_i}$. This dependence, in turn, could be defined by the Pearson correlation coefficient as $r_{\alpha_i y} = \frac{\alpha_i^T y}{\|\alpha_i\|_2 \|y\|_2} = \frac{\|\alpha_i\|_2^2}{\|\alpha_i\|_2 \|y\|_2} = \frac{1}{\|y\|_2} \|X_{I_i}\beta_{I_i}\|_2$ (note that $r_{\alpha_i y}^2$ is also equal to the classical coefficient of determination and gives the percentage of the total variance of y explained by variables from i th group). Another point of view is perceiving the impact on response vector as the impact on its length. Again, this justifies our choice for groups strengths definition, since in considered situation it holds $\|y\|_2 = \sqrt{\sum_{i=1}^m \|X_{I_i}\beta_{I_i}\|_2^2}$. Observe also that our definition is consistent with the classical definition for the single variable selection, since we have $\|X_i\beta_i\|_2 = |\beta_i|$ for ℓ_2 -normalized column X_i .

For $i \in \{1, \dots, m\}$ denote the rank of submatrix X_{I_i} by r_i . Then X_{I_i} could be represented in the form $X_{I_i} = \tilde{X}_i R_i$, where \tilde{X}_i is submatrix composed of r_i linearly independent columns of X_{I_i} . Define $r := r_1 + \dots + r_m$, $\tilde{X} = [\tilde{X}_1, \dots, \tilde{X}_m] \in M(n, r)$, and $\tilde{\beta} = [\beta_{I_1}^T R_1^T, \dots, \beta_{I_m}^T R_m^T]^T \in \mathbb{R}^r$. Now, (2.7) could be rewritten as $y = \sum_{i=1}^m \tilde{X}_{I_i} \tilde{\beta}_{I_i} + z$. Since $\|X_{I_i}\beta_{I_i}\|_2 = \|\tilde{X}_{I_i} \tilde{\beta}_{I_i}\|_2$ for all i , the subsets of relevant groups in both formulations are identical and we end with the statistically equivalent problems. Summarizing, we can always get rid of linear dependencies inside groups and in further part of this thesis we will assume that design (or experiment) matrix satisfies

$$\|X_i\|_2 = 1, \quad X_{I_j} \text{ has linearly independent columns,} \quad i \in \{1, \dots, p\}, \quad j \in \{1, \dots, m\}. \quad (2.8)$$

It is worth to notice that under linear independence of columns inside each group, the conditions $\|X_{I_i}\beta_{I_i}\|_2 > 0$ and $\|\beta_{I_i}\|_2 > 0$ induce the same subset of groups defined as truly relevant. These two approaches give however different hierarchy of groups in terms of effects strengths, since in general case $\|\beta_{I_i}\|_2 > \|\beta_{I_j}\|_2$ does not imply that $\|X_{I_i}\beta_{I_i}\|_2 > \|X_{I_j}\beta_{I_j}\|_2$.

The idea of incorporating groups into penalization method and considering convex optimization problems is not new. Probably the most known approach is the group LASSO (gLASSO) [15]. This extension of LASSO punishes whole groups of variables instead of particular regressors. For fixed tuning parameter, $\lambda_{gL} > 0$, the gLASSO estimate is most frequently (e.g. [16], [17], [18], [19]) defined as the solution to optimization problem

$$\beta^{gL} := \arg \min_b \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^m X_{I_i} b_{I_i} \right\|_2^2 + \sigma \lambda_{gL} \sum_{i=1}^m \sqrt{l_i} \|b_{I_i}\|_2 \right\}, \quad (2.9)$$

for l_i being number of variables in group i . Here, group i is labeled as relevant, if $\|\beta_{I_i}^{gL}\|_2 > 0$. Coefficients $\sqrt{l_i}$'s could be perceived as a additional penalties added to balance different sizes of groups. To explain the reason behind assigning the larger penalty assigned to more numerous groups, suppose that $\beta_{I_j} = 0$ for some j and we want to get $\beta_{I_j}^{gL} = 0$ by using (2.9). If size of this group is growing up, starting from some point it could be more "profitable" (in terms of getting smaller value of objective function) to use many small coefficients in $\beta_{I_i}^{gL}$, which would decrease $\|y - \sum_{i=1}^m X_{I_i} \beta_{I_i}^{gL}\|_2$ while having a nonzero but a very small value of $\|\beta_{I_j}^{gL}\|_2$. To reduce this effect, one should assign larger penalty to larger groups. To construct penalty term, gLASSO is based on a so-called $l_{1,2}$ norm. For a p dimensional vector b , from a definition we have $\|b\|_{1,2} = \sum_{i=1}^m \|b_{I_i}\|_2$, which can be treated as $\|\cdot\|_1$ norm applied to $\llbracket b \rrbracket_I$. Imposing a penalty on whole groups results in the tendency for solution to be sparse at the groups level and the sparsity is determined by the magnitude of the tuning parameter λ_{gL} . If particular group is included in the model, however, then, most often, all corresponding coefficients will be nonzero [17].

In the spirit of our previous remarks about the group effect strength, the more reliable seems to be an idea of imposing penalties on $\|X_{I_i} b_{I_i}\|_2$ rather than $\|b_{I_i}\|_2$, and using the condition $\|X_{I_i} \beta_{I_i}^{gL}\|_2 > 0$ as a group relevance indicator. The alternative version of gLASSO problem, which is built on that concept, was formulated in [19] as

$$\arg \min_b \left\{ \frac{1}{2} \left\| y - \sum_{i=1}^m X_{I_i} b_{I_i} \right\|_2^2 + \sigma \lambda_{gL} \sum_{i=1}^m \sqrt{l_i} \|X_{I_i} b_{I_i}\|_2 \right\}. \quad (2.10)$$

In section 5 we will look at the connections between both formulations by investigating the link between optimization problems of minimizing functions $\frac{1}{2}\|y - Xb\|_2^2 + g(\|b_{I_1}\|_2, \dots, \|b_{I_m}\|_2)$ and $\frac{1}{2}\|y - Xb\|_2^2 + g(\|X_{I_1}b_{I_1}\|_2, \dots, \|X_{I_m}b_{I_m}\|_2)$ for arbitrary function g . In particular, we will discuss the conditions under which both approaches indicate the same groups as relevant.

A natural question is under what conditions gLASSO performs better than the standard LASSO. This issue was investigated by Huang and Zhang [20]. They introduced the idea of strong group sparsity which measures the complexity of signal using the mixture of group and coefficient sparsity. Authors showed that gLASSO is superior to standard LASSO for strongly group-sparse signals. The question about the gLASSO and LASSO comparison was also addressed by Lounici et al. [21]. They showed that gLASSO can achieve an improvement in the prediction and estimation errors as compared to the LASSO. Authors established oracle inequalities for the prediction and ℓ_2 estimation errors and showed that these bounds hold under a restricted eigenvalue condition on the design matrix. Moreover, they proved that the rate of convergence of their upper bounds is optimal up to a logarithmic factor in a minimax sense, for all estimators over a class of group sparse vectors.

The $\ell_{1,2}$ norm was also considered in other penalization methods. The ideas were often motivated by well performing extensions of classical LASSO. Yuan and Lin [16] proposed algorithms for group selection in the context of LARS and the nonnegative garrote. Meier, van de Geer and Bühlmann [22] studied the group LASSO for logistic regression. They showed consistency under certain conditions and proposed a block coordinate descent algorithm applicable in the high-dimensional design. Wei and Huang [23] proposed the adaptive group LASSO, which is a generalization of the adaptive LASSO. They studied the selection property using the group LASSO as the initial estimate and showed that under certain conditions this method is consistent in group selection.

Methods mentioned above produce estimates which are sparse at the group level, not at the level of individual variables. The goal is the selection of response related groups, without the investigation of the sparsity inside such groups, i.e. whether one particular or more of members are relevant. This is not always appropriate for the data and an alternative approach exists in which important groups are selected together with important variables. Such methods are often referred as a bi-level selection [24], [25], [26]. The tendency to producing desired estimates could be achieved for example by replacing $\ell_{1,2}$ by convex combination of $\ell_{1,2}$ and ℓ_1 norms, i.e. one can consider penalty given by $(1 - \alpha)\lambda_{gL} \sum_{i=1}^m \sqrt{l_i} \|b_{I_i}\|_2 + \alpha\lambda_{gL} \|b\|_1$, for tuning parameters $\alpha \in [0, 1]$ and $\lambda_{gL} \geq 0$, which leads directly to sparse-group lasso [18]. The other idea is to construct non-convex optimization problem basing on bridge penalty: $\lambda \sum_{i=1}^m c_i \|b_{I_i}\|_1^\gamma$, for $\gamma \in (0, 1)$ which yields the group bridge estimator [27]. Authors of [27] showed that method has the powerful oracle group selection property, meaning that under certain assumptions it can correctly select important groups with probability converging to one.

3 Notations, mathematical tools and techniques

In this section we will present necessary definitions and theory. First, however, we will recall and introduce some notations, which are often used in further part of thesis.

For given vector $b \in \mathbb{R}^p$, by $b_{(\cdot)} = (b_{(1)}, \dots, b_{(p)})^\top$ we denote vector composed of the ordered coefficients of b , i.e. $b_{(i)}$ is i th largest coefficient of b and it holds $b_{(1)} \geq \dots \geq b_{(p)}$. To denote vector of absolute values, we will use the notation $|b| := (|b_1|, \dots, |b_p|)^\top$. Consequently, $|b|_{(\cdot)}$ is the vector of ordered magnitudes. For $a, b \in \mathbb{R}^p$, the expression $a \prec b$ is equivalent with conditions $a_i < b_i$ for all i and $a \preceq b$ is equivalent with conditions $a_i \leq b_i$ for all i . Vector or matrix with 0 at each entry, will be denoted simply by 0. The dimensions will be always easily readable from the context. The p by p identity matrix is denoted by \mathbf{I}_p .

Let $I = \{I_1, \dots, I_m\}$ be some subsets of $\{1, \dots, p\}$ (not necessarily creating the partition of $\{1, \dots, p\}$), $l_i := |I_i|$ and let $X \in M(n, p)$, where $M(n, p)$ is the space of n by p matrices. By $\llbracket b \rrbracket_I$

and $\llbracket b \rrbracket_{X,I}$ we will denote m dimensional vectors, defined as

$$\llbracket b \rrbracket_I := (\|b_{I_1}\|_2, \dots, \|b_{I_m}\|_2)^\top, \quad \llbracket b \rrbracket_{X,I} := (\|X_{I_1} b_{I_1}\|_2, \dots, \|X_{I_m} b_{I_m}\|_2)^\top. \quad (3.1)$$

Especially important in further part of the thesis is situation when columns in X are orthogonal, i.e. $X_i^\top X_j = 0$. We refer to such a case as *orthogonal situation* and we will say that X is orthogonal if and only if $X^\top X = \mathbf{I}_p$. Notice, that this is not fully consistent with standard definition, in which also condition $XX^\top = \mathbf{I}_p$ is assumed to be satisfied and then X has to be square matrix. Our convention implies only that $p \leq n$. The submatrix of X created from columns with indices from I_i will be denoted as X_{I_i} . We assume that this notation has always priority over operations applied to matrix, e. g. $X_{I_i}^\top$ is the transposition of submatrix, not the submatrix of transposition. The i th column of X will be simply denoted as X_i . Similarly, X_{I_i, I_i} is a submatrix of X formed with entries of rows and columns indexed by I_i .

We will refer to the situation: $X_{I_i}^\top X_{I_i} = \mathbf{I}_{l_i}$, for $i \in \{1, \dots, m\}$, by saying that X is orthogonal inside groups (columns from different groups do not have to be orthogonal). The next important case is when columns coming from different groups are orthogonal (with possibly nonzero inner products of columns in the same group), i.e. $X_{I_i}^\top X_{I_j} = 0$ for all $i \neq j$. If this is the case, we will say that X is orthogonal at groups level. Clearly, if X is orthogonal both, inside groups and at groups level, then X is orthogonal. Since columns in design matrices are most often centered to zero mean and normalized to unit ℓ_2 norm, we say about the strength of correlation between X_i and X_j referring to the value $|X_i^\top X_j|$.

We are starting the overview of the mathematical tools from recalling the important rearrangement inequality. In the next subsections we will focus on results connected with convexity and duality theory, such as conjugate functions, dual norms and dual optimization problems. We are basing mainly on the extensive book *Convex Optimization* of S. Boyd and L. Vandenberghe [28].

3.1 Permutations and rearrangement inequality

Definition 1. Suppose that some permutation $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ is given. We will say that matrix P_π is permutation matrix corresponding to permutation π , when P_π is of form $P_\pi = (e_{\pi(1)}, \dots, e_{\pi(p)})^\top$, where e_i is column vector with one in the i th and zero in every other position.

Proposition 1. For any permutation P_π corresponding to permutation π :

- $P_\pi x = (x_{\pi(1)}, \dots, x_{\pi(p)})^\top$,
- $P_\pi e_i = e_{\pi^{-1}(i)}$,
- $P_{\pi^{-1}} = P_\pi^{-1} = P_\pi^\top$.

Theorem 2 (Rearrangement inequality, [29]). Let $\{x_i\}_{i=1}^p$ and $\{y_i\}_{i=1}^p$ be ordered sequences, i.e.

$$x_1 \geq \dots \geq x_p, \quad y_1 \geq \dots \geq y_p$$

and $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ be any permutation. Then

$$\sum_{i=1}^p x_i y_i \geq \sum_{i=1}^p x_{\pi(i)} y_i \geq \sum_{i=1}^p x_{p-i+1} y_i. \quad (3.2)$$

3.2 Optimization problems and duality

Let f_0, f_1, \dots, f_k and h_1, \dots, h_s be functions defined on \mathbb{R}^p . We consider optimization problem

$$\begin{aligned} & \underset{b}{\text{minimize}} && f_0(b) \\ & \text{subject to} && f_i(b) \leq 0, \quad i = 1, \dots, k \cdot \\ & && h_i(b) = 0, \quad i = 1, \dots, s \end{aligned} \quad (3.3)$$

We say that $b \in \mathbb{R}^p$ is the *optimization variable* and the function $f_0 : \mathbb{R}^p \rightarrow \mathbb{R}$ the *objective function* or *cost function*. The inequalities $f_i(b) \leq 0$ are called *inequality constraints*, and the equations $h_i(b) = 0$ are called the *equality constraints*. If there are no constraints (i.e., $k = s = 0$), we say the problem (3.3) is *unconstrained*.

Definition 2. We are saying that b is *feasible* when it satisfies the constraints $f_i(b) \leq 0$, $i = 1, \dots, k$ and $h_i(b) = 0$, $i = 1, \dots, s$. If b is feasible and additionally for $i = 1, \dots, k$ it holds $f_i(b) < 0$, we say that b is *strictly feasible*. Vector b^* is *solution* to (3.3), if b^* is feasible and for all feasible $b \in \mathbb{R}^p$ it holds $f_0(b) \geq f_0(b^*)$. We will denote the set of all solutions by B^* . We are saying that problem has a *unique solution*, if B^* is a *singleton*.

Definition 3. Suppose that B^* is nonempty. Then, for all vectors from B^* , the objective function takes the same value, which we will denote by f_0^* and call the *optimal value* of f_0 . If B^* is empty, then we put $f_0^* := -\infty$.

Definition 4. We are saying that function $L : \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}^s \rightarrow \mathbb{R}$ is *Lagrangian* associated with the problem (3.3), if L is of form

$$L(b, \nu, \mu) = f_0(b) + \sum_{i=1}^k \nu_i f_i(b) + \sum_{i=1}^s \mu_i h_i(b). \quad (3.4)$$

We refer to μ_i as *Lagrange multiplier* associated with the i th equality constraint and to ν_i as *Lagrange multiplier* associated with the i th inequality constraint. Variables μ and ν are called the *dual variables* associated with the problem (3.3).

Definition 5. The *Lagrange dual function*, $g : \mathbb{R}^k \times \mathbb{R}^s \rightarrow \mathbb{R} \cup \{-\infty\}$, is defined as

$$g(\nu, \mu) = \inf_b L(b, \nu, \mu) = \inf_b \left\{ f_0(b) + \sum_{i=1}^k \nu_i f_i(b) + \sum_{i=1}^s \mu_i h_i(b) \right\}. \quad (3.5)$$

Using this function as an objective, we define the *Lagrange dual problem* as

$$\begin{aligned} & \underset{\nu, \mu}{\text{maximize}} && g(\nu, \mu) \\ & \text{subject to} && \nu \geq 0 \end{aligned} \quad (3.6)$$

We refer to the pair (ν^*, μ^*) , being solution to (3.5), as the *dual solution*. Denote by f^* and g^* the optimal values for primal (3.3) and dual (3.6) problem respectively. Then always $g^* \leq f^*$, which is called *weak duality*. The value $f^* - g^*$ is known as *duality gap* and we say that *strong duality* holds if duality gap is equal to zero.

Definition 6. Suppose that $\|\cdot\|$ is some norm on \mathbb{R}^p . We will use notation $\|\cdot\|^D$ to refer to the associated dual norm, defined as

$$\|x\|^D := \sup_b \{x^\top b : \|b\| \leq 1\}. \quad (3.7)$$

Example 1. Consider $h(b) := \|b\|_2$ and assume that $\|b\|_2 \leq 1$. From Cauchy–Schwarz inequality we have $x^\top b \leq \|x\|_2 \|b\|_2 \leq \|x\|_2$. On the other hand for $x \neq 0$, $x^\top \frac{x}{\|x\|_2} = \|x\|_2$. If $x = 0$, then $\sup_b \{x^\top b : \|b\|_2 \leq 1\} = 0 = \|x\|_2$ and consequently we get that the dual of the ℓ_2 norm is ℓ_2 norm.

Theorem 3. In finite dimensional space the dual of the dual norm is the original norm, i.e.

$$(\|\cdot\|^D)^D = \|\cdot\|.$$

Proof. Let $(X, \|\cdot\|)$ be finite dimensional normed space over \mathbb{R} . Consider its dual normed space X' , that consists of all continuous linear functionals, $f : X \rightarrow \mathbb{R}$, and is equipped with the dual norm $\|\cdot\|'$, defined by $\|f\|' = \sup \{|f(x)| : x \in X, \|x\| \leq 1\}$. If X is finitely dimensional, then $(X', \|\cdot\|')$ corresponds to $(X^D, \|\cdot\|^D)$, where $X^D := \{f_b, b \in X\}$, for f_b being function defined as $f_b(x) = b^T x$ for $x \in X$. Therefore, the claim follows immediately from James' theorem [30], which states that a Banach space $(X, \|\cdot\|)$ is reflexive if and only if every continuous linear functional on X attains its maximum on the closed unit ball in $(X, \|\cdot\|)$. ■

Definition 7. For $f : \mathbb{R}^p \rightarrow \mathbb{R}$, the conjugate function $f^* : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{\infty\}$, is a function given by formula

$$f^*(x) := \sup_b \{x^T b - f(b)\}. \quad (3.8)$$

For issues raised in further part important will be theorem concerning the dependence between conjugate and dual norm.

Theorem 4 (Example 3.26 in [28]). Let $\|\cdot\|$ be norm on \mathbb{R}^p , with dual norm $\|\cdot\|^D$. Then, the conjugate of $f(b) = \|b\|$ is given by the formula

$$f^*(x) = \begin{cases} 0, & \|x\|^D \leq 1 \\ \infty, & \text{otherwise} \end{cases}. \quad (3.9)$$

3.3 Convexity

Definition 8. We are saying that $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is a convex function if for all $b, \tilde{b} \in \mathbb{R}^p$ and $0 \leq \alpha \leq 1$, we have

$$f(\alpha b + (1 - \alpha)\tilde{b}) \leq \alpha f(b) + (1 - \alpha)f(\tilde{b}). \quad (3.10)$$

We are saying that f is strictly convex if inequality (3.10) is strict for $b \neq \tilde{b}$ and $0 < \alpha < 1$.

Theorem 5. [Operations that preserve convexity, subsection 3.2 in [28]] Let f_1, \dots, f_m be convex functions. Then

- nonnegative weighted sum of $\{f_i\}_{i=1}^m$, i.e $f := w_1 f_1 + \dots + w_m f_m$, is convex for $w_i \geq 0$;
- pointwise maximum, $f(b) := \max \{f_1(b), \dots, f_m(b)\}$, is convex;
- suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $A \in M(n, p)$, $y \in \mathbb{R}^n$ and define $g : \mathbb{R}^p \rightarrow \mathbb{R}$ by $g(b) = f(Ab + y)$. Then, g is convex function;
- let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be convex function. Then for each $b, h \in \mathbb{R}^p$ function $F : \mathbb{R} \rightarrow \mathbb{R}$, defined as $F(t) := f(b + th)$, is convex.

Definition 9. We say that problem of form (3.3) is a convex optimization problem, if f_0, f_1, \dots, f_k are convex and h_1, \dots, h_s are affine functions.

Proposition 2. Suppose that the objective function in convex optimization problem is strictly convex. If there exists solution, then this solution is unique.

Theorem 6 (Subsection 5.3.2 in [28]). Consider convex optimization problem of form (3.3) and suppose that there exists strictly feasible vector $b \in \mathbb{R}^p$. Then strong duality holds and $f^* = g^*$.

Theorem 6 is of practical importance, since it enables to construct proper stopping criteria in iterative algorithms designed to find numerical solution. We will also use this result in such context.

3.4 Subgradients and KKT conditions

Suppose that f is convex and differentiable function, $f : \mathbb{R}^p \rightarrow \mathbb{R}$. Then for all $b, h \in \mathbb{R}^p$ it holds $f(b+h) \geq f(b) + (\nabla f(b))^\top h$. This inequality was used to extend the concept of gradients for wider class of functions, yielding important tools in convex analysis. In this subsection we will recall useful definitions and results.

Definition 10. Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$. We say that g is a subgradient of f at $b \in \mathbb{R}^p$ if

$$f(b+h) \geq f(b) + g^\top h, \text{ for all } h \in \mathbb{R}^p.$$

The set of all subgradients of f at b is called the subdifferential of f at b and is denoted by $\partial f(b)$. We say that f is subdifferentiable if the subdifferential of f is nonempty, for all $b \in \mathbb{R}^p$.

Definition 11. We say that a function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is closed if for each $\alpha \in \mathbb{R}$, the sublevel set $\{x \in \mathbb{R}^p : f(x) \leq \alpha\}$ is a closed set.

Theorem 7 (Theorem 3.1.13 in [31]). Suppose that $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is closed convex function. Then, for each $b \in \mathbb{R}^p$, the subgradient $\partial f(b)$ is a nonempty bounded set (in particular f is subdifferentiable).

Suppose now that f is convex function and for some $b, g \in \mathbb{R}^p$ it occurs $f(b+h) \geq f(b) + g^\top h$, for $h \in H$, where H is open set containing zero. Let $h_0 \in \mathbb{R}^p$ be arbitrary vector. From Theorem 5, function $F : \mathbb{R} \rightarrow \mathbb{R}$, defined as $F(t) := f(b+th_0) - tg^\top h_0$, is convex. There exists $t_0 \in (0, 1)$ such that $t_0 h_0 \in H$, what gives

$$f(b) \leq F(t_0) = F((1-t_0) \cdot 0 + t_0 \cdot 1) \leq (1-t_0)f(b) + t_0 F(1) \quad (3.11)$$

and $f(b+h_0) \geq f(b) + g^\top h_0$ as a result. Above reasoning leads to

Corollary 1. For any open set H containing zero the subdifferential of convex function f at b could be equivalently defined as set of vectors g satisfying $f(b+h) \geq f(b) + g^\top h$, for all $h \in H$.

Proposition 3 (Properties of subgradients, Lemmas 3.1.7, 3.1.8 and 3.1.9 in [31]). Suppose that $f : \mathbb{R}^p \rightarrow \mathbb{R}, \psi : \mathbb{R}^p \rightarrow \mathbb{R}, \varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ are convex and closed functions.

- If f is differentiable at b , then $\partial f(b) = \{\nabla f(b)\}$.
- If $f(b) = \psi(Ab + y)$, then $\partial f(b) = A^\top \partial \psi(Ab + y) := \{A^\top g : g \in \partial \psi(Ab + y)\}$.
- If $f(b) = \psi(b) + \varphi(b)$, then $\partial f(b) = \partial \psi(b) + \partial \varphi(b)$, where the last plus sign denotes the algebraic sum of sets.

Proposition 4. Let I_1 and I_2 be nonempty, disjoint subsets of $\{1, \dots, p\}$, $|I_1| = l_1$, $|I_2| = l_2$ and $f(b) = \psi(b_{I_1}) + \varphi(b_{I_2})$, for subdifferentiable functions $f : \mathbb{R}^p \rightarrow \mathbb{R}, \psi : \mathbb{R}^{l_1} \rightarrow \mathbb{R}, \varphi : \mathbb{R}^{l_2} \rightarrow \mathbb{R}$. If $a \in \partial f(b)$, then $a_{I_1} \in \partial \psi(b_{I_1})$ and $a_{I_2} \in \partial \varphi(b_{I_2})$.

Proof. Take any $h \in \mathbb{R}^{l_1}$ and define $\tilde{h} \in \mathbb{R}^p$ by conditions $\tilde{h}_{I_1} = h$ and $\tilde{h}_i = 0$ for $i \notin I_1$. Since f is subdifferentiable, we have

$$\psi(b_{I_1} + h) = f(b + \tilde{h}) - \varphi(b_{I_2}) \geq f(b) + a^\top \tilde{h} - \varphi(b_{I_2}) = \psi(b_{I_1}) + a_{I_1}^\top h, \quad (3.12)$$

which gives $a_{I_1} \in \partial \psi(b_{I_1})$. In the same way we can show that $a_{I_2} \in \partial \varphi(b_{I_2})$. ■

Theorem 8. Let $f_\alpha : \mathbb{R}^p \rightarrow \mathbb{R}$ be convex functions for $\alpha \in A$ and let b_0 be any point in \mathbb{R}^p . Define function $f : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{\infty\}$ as

$$f(b) := \sup_{\alpha \in A} f_\alpha(b).$$

If A is compact (in some metric) and the function $\alpha \rightarrow f_\alpha(b)$ is continuous for each b , then

$$\partial f(b_0) = \text{conv} \bigcup \left\{ \partial f_\alpha(b_0) : f_\alpha(b_0) = f(b_0) \right\},$$

where $\text{conv}A$ denotes the convex hull of a set A .

Consider now optimization problem of form (3.3) with subdifferentiable functions f_0, \dots, f_k and h_1, \dots, h_s . We will now recall Karush-Kuhn-Tucker (KKT) conditions for solution.

Definition 12 (KKT conditions). *We say that $(b, \nu, \mu) \in \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}^s$ satisfies the KKT conditions for problem (3.3), if*

$$\begin{aligned} 0 \in \partial f_0(b) + \sum_{i=1}^k \nu_i \partial f_i(b) + \sum_{i=1}^s \mu_i \partial h_i(b) & \quad \text{(Stationary)} \\ \nu_i f_i(b) = 0 \text{ for each } i \in \{1, \dots, k\} & \quad \text{(Complementary slackness)} \\ f_i(b) \leq 0 \text{ and } h_j(b) = 0, \text{ for } i = 1, \dots, k, j = 1, \dots, s & \quad \text{(Primal feasibility)} \\ \nu \succeq 0 & \quad \text{(Dual feasibility)} \end{aligned}$$

Theorem 9. *For a convex optimization problem of form (3.3) with strong duality, it holds*

$$\begin{aligned} & b^* \text{ is primal solution and } (\nu^*, \mu^*) \text{ is dual solution} \\ \iff & (b^*, \nu^*, \mu^*) \text{ satisfy KKT conditions.} \end{aligned}$$

Theorem 10. *Consider unconstrained convex minimization problem with subdifferentiable objective function f . Then b minimizes f if and only if $0 \in \partial f(b)$.*

3.5 Proximal gradient method

Consider unconstrained optimization problem of form

$$\underset{b}{\text{minimize}} \quad f(b) = g(b) + h(b), \quad (3.13)$$

where g and h are convex functions and g is differentiable (for example LASSO and SLOPE are of such a form). There exist efficient methods, namely *proximal gradient algorithms*, which could be applied to find numerical solution for such objective functions. To design efficient algorithms, however, h must be prox-capable, meaning that there exists efficient algorithm for computing the proximal operator for h ,

$$\text{prox}_{th}(y) := \arg \min_b \left\{ \frac{1}{2t} \|y - b\|_2^2 + h(b) \right\}, \quad (3.14)$$

for each $y \in \mathbb{R}^p$ and $t > 0$. The iterative algorithm work as follows. Suppose that in k step $b^{(k)}$ is the current guess. Then, guess $b^{(k+1)}$ is given by

$$b^{(k+1)} := \arg \min_b \left\{ g(b^{(k)}) + \langle \nabla g(b^{(k)}), b - b^{(k)} \rangle + \frac{1}{2t} \|b - b^{(k)}\|_2^2 + h(b) \right\}. \quad (3.15)$$

The two first terms in objective function in (3.15) are Taylor approximation of g , third addend is a proximity term which is responsible for searching an update reasonably close and t can be treated as a step size.

Problem (3.15) could be reformulated to

$$b^{(k+1)} := \arg \min_b \left\{ \frac{1}{2} \|b^{(k)} - t \nabla g(b^{(k)}) - b\|_2^2 + th(b) \right\}, \quad (3.16)$$

hence $b^{(k+1)} = \text{prox}_{th}(b^{(k)} - t \nabla g(b^{(k)}))$, which justifies the need for existence of fast algorithm computing values of proximal operator. In each step the value of t could be changed forming the sequence $\{t_i\}_{i=1}^\infty$. In situation when $g(b) = \frac{1}{2} \|y - Xb\|_2^2$, we get the Procedure 3, given below.

There are many ways to select t_i 's such that the convergence of $f(b^{(k)})$ to the optimal value is theoretically ensured [32], [33].

Procedure 3 Proximal gradient algorithm

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

3.6 Matrix norms and random matrices

Definition 13. Let A be real-valued, $n \times p$ matrix with entries denoted by a_{ij} . We will use the term Frobenius norm to refer to the norm defined as

$$\|A\|_F := \sqrt{\left(\sum_{i=1}^n \sum_{j=1}^p |a_{ij}|^2 \right)}$$

and the term nuclear norm (trace norm) to refer to

$$\|A\|_* := \text{trace} \left(\sqrt{A^\top A} \right), \quad (3.17)$$

where $\sqrt{A^\top A}$ denotes a positive semi-definite matrix B such that $BB = A^\top A$.

Theorem 11. Let A be real-valued $n \times p$ matrix. Then:

- i) the Frobenius norm is equal to $\|A\|_F = \sqrt{\text{trace}(A^\top A)}$;
- ii) if $n = p$, A is symmetric and positive semi-definite, then $\|A\|_* = \text{trace}(A)$,
- iii) it holds $\|A\|_F \leq \|A\|_* \leq \sqrt{r} \|A\|_F$ for r being the rank of A .

Proposition 5. Suppose that random matrices $X \in M(n, p)$ and $A \in M(p, p)$ have expected values. If entries of X and A are independent random variables, then $\mathbb{E}(XAX^\top) = \mathbb{E}(X\mathbb{E}(A)X^\top)$.

Proof. Since $(XAX^\top)_{i,j} = \sum_{k=1}^p x_{i,k}(AX^\top)_{k,j} = \sum_{k=1}^p \sum_{l=1}^p x_{i,k}a_{k,l}x_{j,l}$, it holds

$$\begin{aligned} (\mathbb{E}(XAX^\top))_{i,j} &= \mathbb{E}((XAX^\top)_{i,j}) = \mathbb{E} \left(\sum_{k=1}^p \sum_{l=1}^p x_{i,k}a_{k,l}x_{j,l} \right) = \sum_{k=1}^p \sum_{l=1}^p \mathbb{E}(a_{k,l})\mathbb{E}(x_{i,k}x_{j,l}) = \\ &= \mathbb{E} \left(\sum_{k=1}^p \sum_{l=1}^p x_{i,k}(\mathbb{E}(A))_{k,l}x_{j,l} \right) = \left(\mathbb{E}(X\mathbb{E}(A)X^\top) \right)_{i,j}. \end{aligned} \quad (3.18)$$

■

Definition 14. Suppose X is an $n \times p$ matrix, each row of which is independent realization of a p -variate normal distribution $\mathcal{N}_p(0, \Psi)$ with covariance matrix Ψ . Then the Wishart distribution is the probability distribution of the $p \times p$ random matrix $X^\top X$. One indicates that $X^\top X$ has Wishart distribution by writing $X^\top X \sim \mathcal{W}_p(\Psi, n)$. When $n \geq p$, we will say that $(X^\top X)^{-1}$ follows an inverse Wishart distribution and we will use notation $(X^\top X)^{-1} \sim \mathcal{W}_p^{-1}(\Psi^{-1}, n)$.

Theorem 12. Suppose that $(X^\top X)^{-1} \sim \mathcal{W}_p^{-1}(\Psi^{-1}, n)$ and $n > p + 1$. Then, the expected value of $(X^\top X)^{-1}$ could be expressed as

$$\mathbb{E}[(X^\top X)^{-1}] = \frac{\Psi^{-1}}{n - p - 1}.$$

4 J_λ norm and SLOPE properties

4.1 J_λ norm properties

For nonnegative, nonincreasing sequence $\lambda_1 \geq \dots \geq \lambda_p \geq 0$, consider function $\mathbb{R}^p \ni b \mapsto J_\lambda(b) \in \mathbb{R}$ given by

$$J_\lambda(b) = \sum_{i=1}^p \lambda_i \cdot |b|_{(i)}, \quad (4.1)$$

where $|b|_{(i)}$ is the i th largest absolute value of the elements of b .

Proposition 6. *If $a, b \in \mathbb{R}^p$ are such that $|a| \preceq |b|$, then $|a|_{(\cdot)} \preceq |b|_{(\cdot)}$.*

Proof. Without loss of generality we can assume that a and b are nonnegative and that it holds $a_1 \geq \dots \geq a_p$. We will show that $a_k \leq b_{(k)}$ for $k \in \{1, \dots, p\}$. For this purpose let's fix k and consider the set $S_k := \{b_i : b_i \geq a_k\}$. It is enough to show that S_k contains at least k elements. For each $j \in \{1, \dots, k\}$ we have

$$b_j \geq a_j \geq a_k \implies b_j \in S_k,$$

what proves the last statement. ■

Corollary 2. *Let $a \in \mathbb{R}^p, b \in \mathbb{R}^p$ and $|a| \preceq |b|$ then Proposition (6) instantly gives that $J_\lambda(a) \leq J_\lambda(b)$, since $J_\lambda(a) = \lambda^\top |a|_{(\cdot)} \leq \lambda^\top |b|_{(\cdot)} = J_\lambda(b)$.*

Proposition 7 (Pulling to zero). *For fixed sequence $\lambda_1 \geq \dots \geq \lambda_p \geq 0$, let $b \in \mathbb{R}^p$ be such that $b_j > 0$ for some $j \in \{1, \dots, p\}$. For $\varepsilon \in (0, b_j]$ define:*

$$(b_\varepsilon)_i := \begin{cases} b_j - \varepsilon, & i = j \\ b_i, & \text{otherwise} \end{cases}.$$

Then:

- (i) $J_\lambda(b_\varepsilon) \leq J_\lambda(b)$,
- (ii) if $\lambda \succ 0$, then $J_\lambda(b_\varepsilon) < J_\lambda(b)$.

Proof. Let $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ be a permutation such as $\sum_{i=1}^p \lambda_i (b_\varepsilon)_{(i)} = \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_i$. From Theorem 2

$$J_\lambda(b) - J_\lambda(b_\varepsilon) = \sum_{i=1}^p \lambda_i b_{(i)} - \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_{(i)} \geq \sum_{i=1}^p \lambda_{\pi(i)} b_i - \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_i = \varepsilon \lambda_{\pi(j)} \geq 0. \quad (4.2)$$

If $\lambda \succ 0$, then the last inequality is strict. ■

Proposition 8 (Pulling to the mean). *For fixed sequence $\lambda_1 \geq \dots \geq \lambda_p \geq 0$, let $b \in \mathbb{R}^p$ be such that $b \succeq 0$ and $b_j > b_l$ for some $j, l \in \{1, \dots, p\}$. For $0 < \varepsilon \leq \frac{b_j - b_l}{2}$ define:*

$$(b_\varepsilon)_i := \begin{cases} b_l + \varepsilon, & i = l \\ b_j - \varepsilon, & i = j \\ b_i, & \text{otherwise} \end{cases}.$$

Then:

- (i) $J_\lambda(b_\varepsilon) \leq J_\lambda(b)$,
- (ii) if $\lambda_1 > \dots > \lambda_p$, then $J_\lambda(b_\varepsilon) < J_\lambda(b)$.

Proof. Let $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ be permutation such as $\sum_{i=1}^p \lambda_i (b_\varepsilon)_{(i)} = \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_i$ and $\lambda_{\pi(j)} \geq \lambda_{\pi(l)}$ (this has to be assumed for $\varepsilon = \frac{b_j - b_l}{2}$, otherwise this inequality is guaranteed). From Theorem 2

$$\begin{aligned} J_\lambda(b) - J_\lambda(b_\varepsilon) &= \sum_{i=1}^p \lambda_i b_{(i)} - \sum_{i=1}^p \lambda_i (b_\varepsilon)_{(i)} = \sum_{i=1}^p \lambda_i b_{(i)} - \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_i \geq \\ &\sum_{i=1}^p \lambda_{\pi(i)} b_i - \sum_{i=1}^p \lambda_{\pi(i)} (b_\varepsilon)_i = \varepsilon (\lambda_{\pi(j)} - \lambda_{\pi(l)}) \geq 0. \end{aligned} \quad (4.3)$$

If $\lambda_1 > \dots > \lambda_p$, then the last inequality is strict. \blacksquare

Let now $\lambda \neq 0$ be fixed nonnegative and nonincreasing sequence. It was shown in [2], that sorted ℓ_1 norm could be expressed as $J_\lambda(b) = \sup_x \{b^\top x : x \in C_\lambda\}$, where C_λ is defined as $C_\lambda := \left\{x \in \mathbb{R}^p : \sum_{i=1}^k |x|_{(i)} \leq \sum_{i=1}^k \lambda_i, k = 1, \dots, p\right\}$. We can quickly find the dual norm to J_λ by applying the above property.

Proposition 9. *The dual norm to J_λ is given by $J_\lambda^D(x) := \max \left\{ \frac{|x|_{(1)}}{\lambda_1}, \dots, \frac{\sum_{i=1}^p |x|_{(i)}}{\sum_{i=1}^p \lambda_i} \right\}$.*

Proof. Since $\lambda > 0$, we have $\sum_{i=1}^k \lambda_i \neq 0$ for each $k \in \{1, \dots, p\}$, hence J_λ^D is well defined. We will show that J_λ^D is a norm. It is obvious that $J_\lambda^D(x) = 0$ if and only if $x = 0$ and that $J_\lambda^D(\alpha x) = |\alpha| J_\lambda^D(x)$ for $\alpha \in \mathbb{R}$. Thus it remains to prove that J_λ^D satisfies the triangle inequality. For $k \in \{1, \dots, p\}$, define λ^k by putting $\lambda_j^k := (\sum_{i=1}^k \lambda_i)^{-1}$ for $j \leq k$ and $\lambda_j^k := 0$ for $j > k$. Then for any $x, y \in \mathbb{R}^p$

$$\begin{aligned} J_\lambda^D(x+y) &= \max_k \left\{ \sum_{i=1}^p \lambda_i^k |x+y|_{(i)} \right\} = \max_k \{J_{\lambda^k}(x+y)\} \leq \max_k \{J_{\lambda^k}(x) + J_{\lambda^k}(y)\} \leq \\ &\max_k \{J_{\lambda^k}(x)\} + \max_k \{J_{\lambda^k}(y)\} = J_\lambda^D(x) + J_\lambda^D(y). \end{aligned} \quad (4.4)$$

Clearly, $C_\lambda = \{b : J_\lambda^D(b) \leq 1\}$. Therefore J_λ is dual norm to J_λ^D and the claim follows from fact that the dual of the dual norm is the original norm. \blacksquare

Proposition 10. *The conjugate function for J_λ is function J_λ^* expressed as*

$$J_\lambda^*(x) = \begin{cases} 0, & x \in C_\lambda \\ \infty, & \text{otherwise} \end{cases}. \quad (4.5)$$

Proof. Claim immediately results from Theorem 4 and Proposition 9. \blacksquare

4.2 SLOPE with diagonal experiment matrix

Let $y \in \mathbb{R}^p$ be a fixed vector and d_1, \dots, d_p be positive numbers. We will use notation $\text{diag}(d_1, \dots, d_p)$ to define the diagonal matrix D such as $D_{i,i} = d_i$ for $i = 1, \dots, p$. In this subsection we will present some properties of solution to SLOPE optimization problem with diagonal experiment matrix. Denote $d := (d_1, \dots, d_p)^\top$ and let b^* be the solution to

$$\underset{b}{\text{minimize}} f(b) := \frac{1}{2} \|y - Db\|_2^2 + J_\lambda(b). \quad (4.6)$$

Since f is strictly convex function, the solution to (4.6), b^* , is unique. It is easy to observe, that changing sign of y_i corresponds to changing sign at i th coefficient of solution as well as permuting y_i 's together with d_i 's permutes coefficients of b^* . We collect these properties in proposition below.

Proposition 11. Let $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ be a given permutation with corresponding matrix P_π . Then:

1. $P_\pi D P_\pi^\top = \text{diag}(d_{\pi(1)}, \dots, d_{\pi(p)})$;
2. $b_\pi := P_\pi b^*$ is solution to

$$\underset{b}{\text{minimize}} \quad f_\pi(b) := \frac{1}{2} \left\| P_\pi y - P_\pi D P_\pi^\top b \right\|_2^2 + J_\lambda(b); \quad (4.7)$$

3. If S is a diagonal matrix such as $|S_{ii}| = 1$, then $b_S := S b^*$ is solution to

$$\underset{b}{\text{minimize}} \quad f_S(b) := \frac{1}{2} \left\| S y - D b \right\|_2^2 + J_\lambda(b). \quad (4.8)$$

Proof. To prove 1 it is enough to show that

$$e_j^\top (P_\pi D P_\pi^\top) e_i = (P_\pi^{-1} e_j)^\top D P_\pi^{-1} e_i = e_{\pi(j)}^\top D e_{\pi(i)} = \begin{cases} d_{\pi(i)} & , i = j \\ 0 & , i \neq j \end{cases}. \quad (4.9)$$

To prove 2 suppose that for some \tilde{b} we have $f_\pi(\tilde{b}) < f_\pi(b_\pi)$. Then

$$\frac{1}{2} \left\| P_\pi y - P_\pi D P_\pi^\top \tilde{b} \right\|_2^2 + J_\lambda(\tilde{b}) < \frac{1}{2} \left\| P_\pi y - P_\pi D P_\pi^\top P_\pi b^* \right\|_2^2 + J_\lambda(P_\pi b^*).$$

This yields

$$\frac{1}{2} \left\| y - D(P_\pi^\top \tilde{b}) \right\|_2^2 + J_\lambda(P_\pi^\top \tilde{b}) < \frac{1}{2} \left\| y - D b^* \right\|_2^2 + J_\lambda(b^*),$$

which contradicts the optimality of b^* . The last claim follows simply from the analogical reasoning. \blacksquare

Proposition 12. If $y \succeq 0$, then $b^* \succeq 0$.

Proof. Suppose that for some r it occurs $b_r < 0$ for any $b \in \mathbb{R}^p$. If $y_r = 0$, then taking \hat{b} defined as $\hat{b}_i := \begin{cases} 0, & i = r \\ b_i, & \text{otherwise} \end{cases}$, we get $|\hat{b}| \leq |b|$ and Corollary 2 gives $J_\lambda(\hat{b}) \leq J_\lambda(b)$. Consequently,

$$f(b) - f(\hat{b}) \geq \frac{1}{2} \left\| y - D b \right\|_2^2 - \frac{1}{2} \left\| y - D \hat{b} \right\|_2^2 = \frac{1}{2} (y_r - d_r b_r)^2 - \frac{1}{2} (y_r + d_r \hat{b}_r)^2 = \frac{1}{2} d_r^2 b_r^2 > 0.$$

Hence b could not be the solution. Now consider case when $y_r > 0$ and define \hat{b} by putting $\hat{b}_i := \begin{cases} -b_r, & i = r \\ b_i, & \text{otherwise} \end{cases}$. Then we have $J_\lambda(b) = J_\lambda(\hat{b})$ and

$$f(b) - f(\hat{b}) = \frac{1}{2} (y_r - d_r b_r)^2 - \frac{1}{2} (y_r + d_r b_r)^2 = -2y_r d_r b_r > 0.$$

and, as before, b could not be optimal. \blacksquare

Proposition 13. Let b^* be the solution to problem (4.6), $\{y_i\}_{i=1}^p$ be nonnegative sequence, $\{d_i\}_{i=1}^p$ be the sequence of positive numbers and assume that

$$d_1 y_1 \geq \dots \geq d_p y_p. \quad (4.10)$$

If b^* has exactly r nonzero entries for $r > 0$, then the set $\{1, \dots, r\}$ is the support of b^* .

Proof. It is enough to show that

$$(j \in \{2, \dots, m\}, b_j^* \neq 0) \implies b_{j-1}^* \neq 0.$$

Suppose that this is not true. From Proposition 12 we know that b^* is nonnegative, hence we can find i from $\{2, \dots, m\}$ such as $b_j^* > 0$ and $b_{j-1}^* = 0$. For $\varepsilon \in (0, b_j^*/2)$ define vector b_ε by putting

$$b_\varepsilon(i) := \begin{cases} \varepsilon, & i = j - 1 \\ b_j^* - \varepsilon, & i = j \\ b_i^*, & \text{otherwise} \end{cases}. \quad (4.11)$$

From Proposition 8 we have that $J_\lambda(b_\varepsilon) \leq J_\lambda(b^*)$, which gives

$$\begin{aligned} f(b^*) - f(b_\varepsilon) &\geq \frac{1}{2}(y_{j-1} - d_{j-1}b_{j-1}^*)^2 + \frac{1}{2}(y_j - d_jb_j^*)^2 \\ &\quad - \frac{1}{2}(y_{j-1} - d_{j-1}b_\varepsilon(j-1))^2 - \frac{1}{2}(y_j - d_jb_\varepsilon(j))^2 = \\ &\quad \varepsilon \left(A - \frac{d_{j-1}^2 + d_j^2}{2} \cdot \varepsilon \right), \end{aligned} \quad (4.12)$$

for $A := (y_{j-1}d_{j-1} - y_jd_j) + d_j^2b_j^* > 0$.

Therefore, we can find $\varepsilon > 0$ such as $f(b^*) > f(b_\varepsilon)$, which contradicts the optimality of b^* . \blacksquare

Consider now problem (4.6) with arbitrary sequence $\{y_i\}_{i=1}^p$. Suppose that b^* has exactly $r > 0$ nonzero coefficients and that $\pi : \{1, \dots, p\} \rightarrow \{1, \dots, p\}$ is permutation which gives the order of magnitudes for Dy , i.e. $d_{\pi(1)}|y|_{\pi(1)} \geq \dots \geq d_{\pi(p)}|y|_{\pi(p)}$. Let P_π be matrix corresponding to π and S be diagonal matrix such as $S_{i,i} = \text{sgn}(y_i)$. Then, from Proposition 11 the solution to problem

$$\arg \min_b \left\{ \frac{1}{2} \|P_\pi S y - P_\pi D P_\pi^\top b\|_2^2 + J_\lambda(b) \right\} \quad (4.13)$$

is given by $P_\pi S b^*$. Moreover, $P_\pi S y$ and $P_\pi D P_\pi^\top$ in (4.13) satisfy assumptions of Proposition 13, which gives important

Corollary 3. *If b^* is the solution to (4.6) having exactly $r > 0$ nonzero coefficients and π is permutation which orders $D|y|$, i.e. $d_{\pi(i)}|y|_{\pi(i)} = |Dy|_{(i)}$ for $i = 1, \dots, p$, then the support of b^* is composed of the set $\{\pi(1), \dots, \pi(r)\}$.*

The next three lemmas were proved in [2] in situation when $d_1 = \dots = d_p = 1$. We will follow the reasoning from this paper to prove the generalized claims. The main difference is that in general case the solution to considered problem (4.6) does not have to be nonincreasingly ordered, under assumption that $d_1 y_1 \geq \dots \geq d_p y_p \geq 0$ (which is the case for $d_1 = \dots = d_p = 1$). This makes that generalizations of proofs presented in [2] are not straightforward.

Lemma 1. *Consider nonnegative sequence $\{y_i\}_{i=1}^p$ and sequence of positive numbers $\{d_i\}_{i=1}^p$ such as $d_1 y_1 \geq \dots \geq d_p y_p$. If b^* is solution to problem (4.6) having exactly r nonzero entries, then for every $j \leq r$ it holds that*

$$\sum_{i=j}^r (d_i y_i - \lambda_i) > 0 \quad (4.14)$$

and for every $j \geq r + 1$

$$\sum_{i=r+1}^j (d_i y_i - \lambda_i) \leq 0. \quad (4.15)$$

The proof of above lemma is presented in the Appendix. We will use this result to prove two next lemmas.

Lemma 2. *Let b^* be solution to problem (4.6) with nonnegative, nonincreasing sequence $\{\lambda_i\}_{i=1}^p$. Let $R(b^*)$ be number of all nonzeros in b^* and $r \geq 1$. Then, for any $i \in \{1, \dots, p\}$*

$$\{y : b_i^* \neq 0 \text{ and } R(b^*) = r\} = \{y : d_i|y_i| > \lambda_r \text{ and } R(b^*) = r\}.$$

Proof. Suppose that b^* has $r > 0$ nonzero coefficients and let π be permutation which orders vector $D|y|$. From Corollary 3 it holds that $\{i : b_i^* \neq 0\} = \{\pi(1), \dots, \pi(r)\}$. Define $\tilde{y} := P_\pi S y$ and $\tilde{D} := P_\pi D P_\pi^\top$, for S being the diagonal matrix such as $S_{i,i} = \text{sgn}(y_i)$. Then $P_\pi|b^*$ is solution to problem

$$\arg \min_b \frac{1}{2} \|\tilde{y} - \tilde{D}b\|_2^2 + J_\lambda(b), \quad (4.16)$$

which satisfies the assumptions of Lemma 1. Taking $j = r$ in (4.14) and $r = j + 1$ in (4.15) we immediately get

$$d_{\pi(r)}|y|_{\pi(r)} > \lambda_r \text{ and } d_{\pi(r+1)}|y|_{\pi(r+1)} \leq \lambda_{r+1}. \quad (4.17)$$

We will now show that $\{y : b_i^* \neq 0 \text{ and } R(b^*) = r\} \subset \{y : d_i|y_i| > \lambda_r \text{ and } R(b^*) = r\}$. Fix $i \in \{1, \dots, p\}$ and suppose that b_i^* is nonzero coefficient. Then $i \in \{\pi(1), \dots, \pi(r)\}$ and therefore $d_i|y_i| \geq d_{\pi(r)}|y|_{\pi(r)} > \lambda_r$, thanks to first inequality from (4.17). To show the second inclusion assume that $d_i|y_i| > \lambda_r$. Then, from the second inequality in (4.17), $d_i|y_i| > \lambda_{r+1} \geq d_{\pi(r+1)}|y|_{\pi(r+1)}$, which gives $i \in \{\pi(1), \dots, \pi(r)\}$. ■

Lemma 3. *For given sequence $\{y_i\}_{i=1}^p$, sequence of positive numbers $\{d_i\}_{i=1}^p$, nonincreasing, nonnegative sequence $\{\lambda_i\}_{i=1}^p$ and fixed $j \in \{1, \dots, p\}$, consider following procedure*

- define $\tilde{y} := (y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_p)^\top$, $\tilde{D} := \text{diag}(d_1, \dots, d_{j-1}, d_{j+1}, \dots, d_p)$, $\tilde{d}_i := \tilde{D}_{i,i}$ for $i = 1, \dots, p - 1$ and $\tilde{\lambda} := (\lambda_2, \dots, \lambda_p)^\top$;
- find $\tilde{b}^* := \arg \min_{b \in \mathbb{R}^{p-1}} \frac{1}{2} \|\tilde{y} - \tilde{D}b\|_2^2 + J_{\tilde{\lambda}}(b)$;
- define $\tilde{R}^j(\tilde{b}^*) := |\{i : \tilde{b}_i^* \neq 0\}|$.

Then for $r \geq 1$ it holds $\{y : d_j|y_j| > \lambda_r \text{ and } R(b^*) = r\} \subset \{y : d_j|y_j| > \lambda_r \text{ and } \tilde{R}^j(\tilde{b}^*) = r - 1\}$.

For the sake of readability, the proof of Lemma 3 is provided in the Appendix. Equipped with above lemmas, we are ready to prove the interesting properties of group SLOPE method, which we introduce in the next section.

5 group SLOPE

5.1 Formulation of the optimization problem

Suppose that $I = \{I_1, \dots, I_m\}$ is some partition of the set $\{1, \dots, p\}$, i.e. I_i 's are nonempty sets, $I_i \cap I_j = \emptyset$ for $i \neq j$ and $\bigcup I_i = \{1, \dots, p\}$. Let l_i be the number of elements in group i for $i = 1, \dots, m$. Moreover, let $X \in M(n, p)$ be given matrix satisfying (2.8). We will consider the linear regression model with m groups of form (2.7). Here, the task is to identify the relevant groups, i.e. the set $\{i : \|X_{I_i}\beta_{I_i}\|_2 > 0\}$ which could be perceived as finding the support of vector $\llbracket \beta \rrbracket_{X,I} := (\|X_{I_1}\beta_{I_1}\|_2, \dots, \|X_{I_m}\beta_{I_m}\|_2)^\top$. Since we work with linear independence inside each group, the task is equivalent to estimating the set $\{i : \|\beta_{I_i}\|_2 > 0\}$, being the support of vector $\llbracket \beta \rrbracket_I := (\|\beta_{I_1}\|_2, \dots, \|\beta_{I_m}\|_2)^\top$.

To estimate the nonzero coefficients of $\llbracket \beta \rrbracket_I$, we will use new penalized method, namely group SLOPE (gSLOPE). For given sequence of nonincreasing, nonnegative tuning parameters $\lambda_1, \dots, \lambda_m$, given sequence of positive weights w_1, \dots, w_m , and design matrix X , the gSLOPE, β^{gs} , is defined as

$$\beta^{\text{gs}} := \arg \min_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \sigma J_\lambda(W \llbracket b \rrbracket_I) \right\}, \quad (5.1)$$

where W is diagonal matrix defined by equations $W_{i,i} := w_i$, for $i = 1, \dots, m$. The estimate of $\llbracket \beta \rrbracket_I$ support is defined by the indices corresponding to nonzeros of $\llbracket \beta^{\text{gs}} \rrbracket_I$.

Remark 1. *At the time of finishing this thesis, we have been informed that the similar formulation of the group SLOPE was proposed in an independent work of Alexej Gossmann et al. [34]. However [34] considers only the case when the weights w_i are equal to the square root of the group size and does not use the orthogonalization inside groups. In comparison to [34], who propose a Monte Carlo approach for estimating the regularizing sequence, this thesis provides theoretically justified choice of the smoothing parameters in orthogonal situation and investigates FDR control in near-orthogonal cases.*

It is easy to see that when one consider p groups containing only one variable (i.e. single-groups situation), then taking all weights equal to one reduces (5.1) to SLOPE (2.4). On the other hand, taking $w_i = \sqrt{l_i}$ and putting $\lambda_1 = \dots = \lambda_m =: \lambda_{gL}$, immediately gives gLASSO problem (2.9) with starting parameter λ_{gL} . Formulation (5.1) could be therefore treated both as the extension to SLOPE and the extension to group LASSO. To show that the objective in problem (5.1) is convex function, we will prove

Proposition 14. *Function $J_{\lambda,W,I}(b) := J_\lambda(W \llbracket b \rrbracket_I)$ is a norm for any nonnegative, nonincreasing sequence $\{\lambda_i\}_{i=1}^m$ containing at least one nonzero element, partition I of the set $\{1, \dots, p\}$ and diagonal matrix W with positive elements on diagonal.*

Proof. It is easy to see that $J_{\lambda,W,I}(b) = 0$ if and only if $b = 0$ and that for any scalar $\alpha \in \mathbb{R}$ it occurs $J_{\lambda,W,I}(\alpha b) = |\alpha| J_{\lambda,W,I}(b)$. We will show that $J_{\lambda,W,I}$ satisfies the triangle inequality. Let a, b be any vectors from \mathbb{R}^p . From the positivity of w_i 's we have $W \llbracket a + b \rrbracket_I \preceq W \llbracket a \rrbracket_I + W \llbracket b \rrbracket_I$. Therefore, Corollary 2 yields

$$\begin{aligned} J_{\lambda,W,I}(a + b) &\leq J_\lambda(W \llbracket a \rrbracket_I + W \llbracket b \rrbracket_I) \leq J_\lambda(W \llbracket a \rrbracket_I) + J_\lambda(W \llbracket b \rrbracket_I) = \\ &J_{\lambda,W,I}(a) + J_{\lambda,W,I}(b), \end{aligned} \quad (5.2)$$

since J_λ is the norm. ■

It is worth to notice, that when $p > n$, the objective in (5.1) is not strictly convex, and optimization problem may not have a unique solution. The typical property of penalized methods defining estimate by minimizing objective of form $\frac{1}{2} \|y - Xb\|_2^2 + g(b)$, with sparsity inducing norm g , is that if the predictor variables are drawn from a continuous probability distribution, then there is a unique solution with probability one, regardless of the sizes of n and p . Such result for LASSO could be for example found in ([35]). As a representative of the same class of penalized method, gSLOPE is likely to share the same type property, however, this issue is beyond of the scope of this thesis.

The reason for the merging of group LASSO approach with SLOPE was to use the J_λ norm property in the context of controlling the fraction of falsely identified groups. We will now provide definitions and notations enabling to introduce this new form of type-one error.

Definition 15. *Consider model (2.7) with design matrix satisfying (2.8) and let $\hat{\beta}$ be the some estimate of β (treated as random vector dependent on Y). We define two random variables: the number of all discovered groups (Rg) and the number of falsely discovered groups (Vg), as*

$$Rg := |\{i : \|\hat{\beta}_{I_i}\|_2 \neq 0\}|, \quad Vg := |\{i : \|\beta_{I_i}\|_2 = 0, \quad \|\hat{\beta}_{I_i}\|_2 \neq 0\}|.$$

Definition 16. We define the false discovery rate for groups (gFDR) as

$$gFDR := \mathbb{E} \left[\frac{Vg}{\max\{Rg, 1\}} \right]. \quad (5.3)$$

Our main goal was to propose the regularization method enabling (in case of relatively weak correlations between variables coming from different groups) to achieve a gFDR control, meaning that for predefined parameter (target gFDR level), $q \in (0, 1)$, particular tuning parameters could explicitly determined in such way that $gFDR \leq q$. In further part of thesis we will show that gSLOPE has very interesting properties in that context. In particular, consider idealistic situation with orthogonal all pairs of columns included to different groups. Under such a design, we will give precise formula for the sequence $\{\lambda_i\}_{i=1}^m$ such that the inequality $gFDR \leq q$ is guaranteed, for estimate given by gSLOPE.

5.2 Standardization of design matrix

Let X be any matrix satisfying (2.8) and $I = \{I_1, \dots, I_m\}$ be some partition of $\{1, \dots, p\}$. For any i , we can consider the decomposition of form $X_{I_i} = U_i R_i$, where $U_i^T U_i = \mathbf{I}_{I_i}$ (such decomposition could be for example obtained from SVD decomposition of X_{I_i}). Define new design matrix, $X_U \in M(n, p)$, by conditions $(X_U)_{I_i} := U_i$ and new vector of interest, $\tilde{\beta} \in \mathbb{R}^p$ by putting $\tilde{\beta}_{I_i} := R_i \beta_{I_i}$, for $i \in \{1, \dots, m\}$. Simply, X_U and $\tilde{\beta}$ defines identical groups effects as in the starting situation, and we one can focus on estimating of $\llbracket \tilde{\beta} \rrbracket_I$ support. Moreover, X_U is orthogonal inside each group, meaning that each submatrix $(X_U)_{I_i}$ consists of orthogonal columns. We will refer to the above procedure as *standardization* of X (in further part of thesis we will just say that X was standardized, without changing the notation).

Standardization is important from both: the theoretical and the practical points of view. It enables to derive results which are independent on the data structure inside groups and provides faster working of numerical algorithms used to solve problems of form (2.9) and (5.1). We want to emphasize, however, that in general situation applying gSLOPE (or gLASSO) to standardized and unstandardized data is not consistent in terms of groups estimated as relevant. In the literature there is much confusion about orthonormalizing within groups [19]. Some authors work with assumption about the orthogonality inside groups, for example [16], while others authors ([18], [20], [20]) make no mention about it. This issue, in more details, was discussed by Noah Simon and Rob Tibshirani [19] in the context of gLASSO. They showed the efficacy of the standardized version over unstandardized on real and simulated data sets.

We will now show that standardization of design matrix is important step which links our definition of group effects sizes with imposing penalties on $\|b_{I_1}\|_2, \dots, \|b_{I_m}\|_2$ while using regularization methods. Let X_U be matrix after (some) standardization, i.e for i it holds $(X_U)_{I_i} = U_i$, where U_i is matrix with orthogonal columns satisfying $U_i R_i = X_{I_i}$ for invertible matrix R_i . Consider optimization problem with the objective $\frac{1}{2} \|y - X_U b\|_2^2 + g(\|b_{I_1}\|_2, \dots, \|b_{I_m}\|_2)$, where g is any function. Moreover, suppose that b^* is unique solution to this problem. Then after substituting $b_{I_i} = R_i c_{I_i}$ for $i \in \{1, \dots, m\}$, we get the equivalent formulation

$$\begin{cases} c^* := \arg \min_c \left\{ \frac{1}{2} \|y - Xc\|_2^2 + g(\|X_{I_1} c_{I_1}\|_2, \dots, \|X_{I_m} c_{I_m}\|_2) \right\}, \\ b_{I_i}^* = R_i c_{I_i}^*, \quad i = 1, \dots, m \end{cases}, \quad (5.4)$$

since $\|R_i c_{I_i}\|_2 = \|U_i R_i c_{I_i}\|_2 = \|X_{I_i} c_{I_i}\|_2$. Now, $\|b_{I_i}^*\|_2 > 0$ if and only if $\|c_{I_i}^*\|_2 > 0$ and we see that estimation of $\llbracket \beta \rrbracket_I$ with standardized data and penalties imposed on $\|b_{I_1}\|_2, \dots, \|b_{I_m}\|_2$ corresponds exactly to penalizing group effects (defined in the introduction) and using original design matrix. Furthermore, when one decides to orthogonalize data inside groups, the chosen decompositions do not have an impact on the subset of groups estimated as relevant, since c^* does not depend on U_i and R_i .

In this thesis we will focus on using gSLOPE with standardized data which is our recommended procedure. We want to highlight, however, that the numerical algorithm which we use to solve (5.1) is not restricted by assumption about orthogonality inside groups and could be applied to general design matrix X .

5.3 Orthogonal situation

Suppose that experiment matrix is orthogonal at groups level, i.e. it holds $X_{I_i}^\top X_{I_j} = 0$, for various $i, j \in \{1, \dots, m\}$. After standardizing data, we get $X_{I_i}^\top X_{I_i} = \mathbf{I}_{I_i}$ for all i and $X^\top X = \mathbf{I}_p$, as a result. If $n = p$, i.e. X is square, orthogonal matrix, we also have $XX^\top = \mathbf{I}_p$ and it occurs $\|X^\top(y - Xb)\|_2^2 = (y - Xb)^\top XX^\top(y - Xb) = \|y - Xb\|_2^2$ for $b \in \mathbb{R}^p$, which enables to simplify the model fitting term in the objective in gSLOPE problem, by putting $\|y - Xb\|_2^2 = \|X^\top y - b\|_2^2$. For the general case with $n \leq p$, we can extend X to square matrix by adding new orthogonal columns and defining $X_C := [X | C]$, where C is composed of vectors (columns) being some complement to orthogonal basis of \mathbb{R}^p . For $y \in \mathbb{R}^n$ and $b \in \mathbb{R}^p$ we get:

$$\|y - Xb\|_2^2 = \|X_C^\top(y - Xb)\|_2^2 = \left\| \begin{bmatrix} X^\top \\ C^\top \end{bmatrix} y - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2^2 = \|X^\top y - b\|_2^2 + \text{const}, \quad (5.5)$$

which implies that under orthogonal situation problem (5.1) could be recast as

$$\arg \min_b \left\{ \frac{1}{2} \|\tilde{y} - b\|_2^2 + \sigma J_\lambda(W\llbracket b \rrbracket_I) \right\}, \quad (5.6)$$

for $\tilde{y} := X^\top y$ being realization of random variable $\tilde{Y} := X^\top Y$. Notice that $\tilde{Y} \sim \mathcal{N}(\beta, \sigma^2 \mathbf{I}_p)$ and we will assume that $\|\tilde{y}_{I_i}\|_2 \neq 0$ for all i . After introducing new variable to problem (5.6), namely $c \in \mathbb{R}^m$, we get the equivalent formulation

$$\arg \min_{b,c} \left\{ \frac{1}{2} \|\tilde{y} - b\|_2^2 + \sigma J_\lambda(c) : c = W\llbracket b \rrbracket_I, c \succeq 0 \right\}. \quad (5.7)$$

Proposition 15. *Let $f(b, c) : \mathbb{R}^p \times \mathbb{R}^m \rightarrow \mathbb{R}$ be any function and consider optimization problem $\arg \min_{b,c} \{f(b, c) : (b, c) \in \mathcal{D}\}$ with unique solution (b^*, c^*) and feasible set $\mathcal{D} \subset \mathbb{R}^p \times \mathbb{R}^m$. Define $\mathcal{D}^c := \{c \in \mathbb{R}^m \mid \exists b \in \mathbb{R}^p : (b, c) \in \mathcal{D}\}$. Suppose that for any $c \in \mathcal{D}^c$, there exists unique solution, b^c , to problem $\arg \min_b \{f(b, c) : (b, c) \in \mathcal{D}\}$. Moreover, assume that also the solution to $\arg \min_c \{f(b^c, c) : c \in \mathcal{D}^c\}$ is unique. Then, it occurs*

$$\begin{cases} c^* = \arg \min_c \{f(b^c, c) : c \in \mathcal{D}^c\} \\ b^* = b^{c^*} \end{cases}. \quad (5.8)$$

Proof. Suppose that there exists $(b^0, c^0) \in \mathcal{D}$, such that $f(b^0, c^0) < f(b^*, c^*)$, where b^* and c^* are defined as in (5.8). We have

$$f(b^{c^0}, c^0) \leq f(b^0, c^0) < f(b^*, c^*) = f(b^{c^*}, c^*), \quad (5.9)$$

which leads to contradiction with definition of c^* . ■

We will apply the above proposition to (5.7). Let (b^*, c^*) be solution to (5.7). Then b^* is also solution to convex problem (5.7) with strictly convex objective function and therefore is unique. Since

$c^* = W\llbracket b^* \rrbracket_I$, c^* is unique as well. In considered situation $\mathcal{D}^c = \{c : c \succeq 0\}$. We will start with finding b^c for any $c \in \mathcal{D}^c$, i.e. we will solve problem $b^c = \arg \min_b \left\{ \frac{1}{2} \|\tilde{y} - b\|_2^2 + \sigma J_\lambda(c) : c = W\llbracket b \rrbracket_I \right\}$. The additive constant in the objective could be omitted. Moreover, for each $i \in \{1, \dots, m\}$ we have

$$b_{I_i}^c = \arg \min_{b_{I_i}} \left\{ \frac{1}{2} \|\tilde{y}_{I_i} - b_{I_i}\|_2^2 : \frac{1}{2} w_i^2 \|b_{I_i}\|_2^2 - \frac{1}{2} c_i^2 = 0 \right\}. \quad (5.10)$$

The principle of Lagrange multipliers yields $b_{I_i}^c = (w_i \|\tilde{y}_{I_i}\|_2)^{-1} c_i \tilde{y}_{I_i}$ or $b_{I_i}^c = -(w_i \|\tilde{y}_{I_i}\|_2)^{-1} c_i \tilde{y}_{I_i}$. Putting these arguments to objective function gives respectively

$$\|\tilde{y}_{I_i} - b_{I_i}^c\|_2^2 = (\|\tilde{y}_{I_i}\|_2 - w_i^{-1} c_i)^2 \quad \text{or} \quad \|\tilde{y}_{I_i} - b_{I_i}^c\|_2^2 = (\|\tilde{y}_{I_i}\|_2 + w_i^{-1} c_i)^2,$$

which means that the former vector corresponds to minimum. From Proposition 15, we get following procedure for solution, $\beta^E gS$, to problem (5.6)

$$\begin{cases} c^* = \arg \min_c \left\{ \frac{1}{2} \sum_{i=1}^m (\|\tilde{y}_{I_i}\|_2 - w_i^{-1} c_i)^2 + J_{\sigma\lambda}(c) \right\} \\ \beta^E gS_{I_i} = c_i^* (w_i \|\tilde{y}_{I_i}\|_2)^{-1} \tilde{y}_{I_i}, \quad i = 1, \dots, m \end{cases} \quad (5.11)$$

(notice that we applied Proposition 12 to omit the constraints $c \succeq 0$ and that c^* is unique from Proposition 2). The above procedure yields conclusion, that indices of groups estimated by gSLOPE as relevant coincide with the support of solution to SLOPE problem with diagonal matrix having inverses of weights w_1, \dots, w_m on diagonal. For further part of reasoning we will need the following definitions.

Definition 17. We will say that random variable X_1 has 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 χ^2 distribution with l degrees of freedom. We will say that random variable X_1 has 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 χ distribution with l degrees of freedom. We will use notation $X_1 \sim \mathcal{S}\chi_l$.

We are now in a position to discuss the issue of gFDR control under orthogonal design. To start with, consider first situation when $l_1 = \dots = l_m =: l$, $w_1 = \dots = w_m =: w$. Moreover, assume that the set $\{1, \dots, m_0\}$ corresponds to indices of zeros in $\llbracket \beta \rrbracket_I$, for some $m_0 \in \{1, \dots, m\}$ (i.e. groups $\{1, \dots, m_0\}$ are truly irrelevant). After multiplying the objective in (5.11) by w^2 and defining $\lambda_{w\sigma} := w^2 \sigma \lambda$, $\Psi_i := w \|\tilde{Y}_{I_i}\|_2$ and $\psi_i := w \|\tilde{y}_{I_i}\|_2$, for $i = 1, \dots, m$, we get the equivalent problem:

$$c^* = \arg \min_c \left\{ \frac{1}{2} \sum_{i=1}^m (\psi_i - c_i)^2 + J_{\lambda_{w\sigma}}(c) \right\}, \quad (5.12)$$

which is exactly the SLOPE problem under orthogonal design. Furthermore, the gFDR for (5.6) coincides with the FDR for (5.12). Notice, that for every $i \leq m_0$, Ψ_i follows scaled χ distribution with scale parameter, $\mathcal{S} = w\sigma$, and l degrees of freedom. Now, analyzing the proof of Theorem 1.1 in [2], one can conclude, that the condition $gFDR \leq q \cdot \frac{m_0}{m}$ is theoretically guaranteed after setting $(\lambda_{w\sigma})_i := F_{w\sigma\chi_l}^{-1} \left(1 - \frac{q \cdot i}{m} \right)$, for $F_{w\sigma\chi_l}$ being the cumulative distribution function of $w\sigma\chi_l$ distribution. This setting immediately yields

$$\lambda_i = F_{\chi_l}^{-1} \left(1 - \frac{q \cdot i}{m} \right), \quad i = 1, \dots, m. \quad (5.13)$$

In Figure 2 we collect results of simulations for gFDR and power (the expected value of the fraction of truly relevant groups which were identified). For each target gFDR level and true support size we generated observations according to (2.7) with $\sigma = 1$ and $m = 1000$ groups, each containing

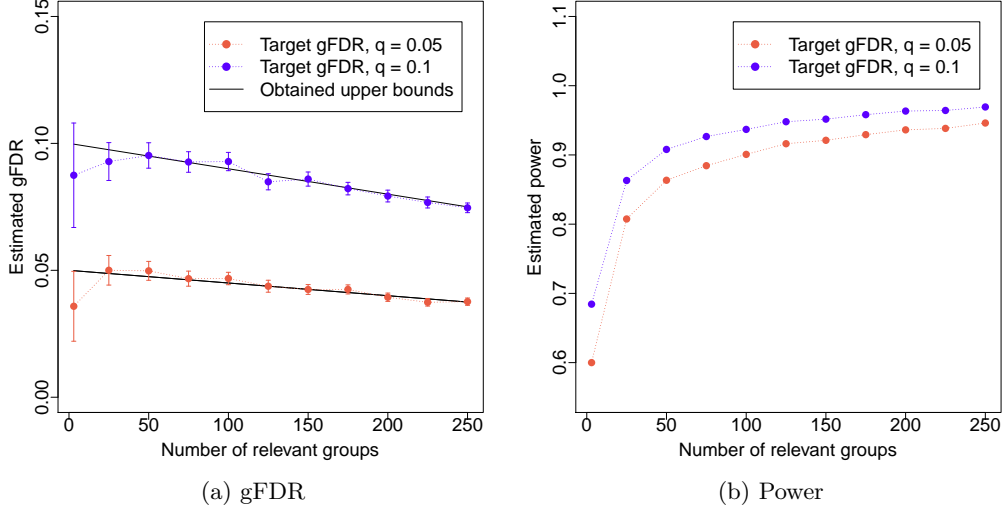


Figure 2: Orthogonal situation with equal groups sizes. For each target gFDR level and true support size, 300 iterations were performed, bars correspond to $\pm 2\text{SE}$. Here, black straight lines are given by $q \cdot ((m - k)/m)$, for k being the true support size.

$l = 5$ variables. We considered orthogonal situation with $n = p = 5000$. For each index i , included to true support, we randomized group effect such as $\|\beta_{I_i}\|_2 := B(m, l)$, for $B(m, l) := \sqrt{4 \log(m)/(1 - m^{-\frac{2}{l}})} - l$. This value is justified in the next subsection.

The main theoretical result in this thesis, concerning gSLOPE, is the procedure for generating tuning parameters for arbitrary groups sizes, $\{l_i\}_{i=1}^m$, and positive weights, $\{w_i\}_{i=1}^m$, which guarantees the control over gFDR when design matrix is orthogonal at groups level, namely Theorem 13. Our result could be treated as the generalization of the Theorem 1.1 from [2]. In proof we will use nontrivial generalizations of lemmas introduced in [2], which were derived in previous section.

Theorem 13 (gFDR control under orthogonal case). *Consider model (2.7) with design matrix X satisfying $X_{I_i}^\top X_{I_j} = 0$, for any $i \neq j$. Denote the number of zero coefficients in $\llbracket \beta \rrbracket_I$ by m_0 . Moreover, fix positive numbers w_1, \dots, w_m and let W be diagonal matrix such as $W_{i,i} = w_i$, for $i = 1, \dots, m$. Apply following steps*

- standardize X ;
- define $\lambda = (\lambda_1, \dots, \lambda_m)^\top$, for $\lambda_i := \max_{j=1, \dots, m} \left\{ \frac{1}{w_j} F_{\chi_{l_j}}^{-1} \left(1 - \frac{q \cdot i}{m} \right) \right\}$, where $F_{\chi_{l_j}}$ is cumulative distribution function of chi distribution with l_j degrees of freedom
- find gSLOPE estimate, $\beta^{\text{gS}} := \arg \min_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \sigma J_\lambda(W \llbracket b \rrbracket_I) \right\}$;

Then, for β^{gS} it holds

$$\text{gFDR} \leq q \cdot \frac{m_0}{m}.$$

Proof. Basing on our previous observation, problem (5.6) is equivalent with (5.11). Define random variables $R := |\{i : c_i^* \neq 0\}|$ and $V := |\{i : \|\beta_{I_i}\|_2 = 0, c_i^* \neq 0\}|$. Clearly, then $Rg = R$ and $Vg = V$. Consequently, it is enough to show that

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

From Corollary 3, applying the same permutation to $\|\tilde{y}_{I_1}\|_2, \dots, \|\tilde{y}_{I_m}\|_2$ and $w_1^{-1}, \dots, w_m^{-1}$ does not have an impact on R and V and without loss of generality we can assume that $\|\beta_{I_1}\|_2, \dots, \|\beta_{I_{m_0}}\|_2$ are truly irrelevant, which gives $\|\beta_{I_1}\|_2 = \dots = \|\beta_{I_{m_0}}\|_2 = 0$ and $\|\beta_{I_j}\|_2 > 0$ for $j > m_0$. Suppose now that r, i are some fixed indices from $\{1, \dots, m\}$. From definition of λ_r ,

$$\lambda_r \geq \frac{1}{w_i} F_{\chi_{i_i}}^{-1} \left(1 - \frac{qr}{m} \right) \implies 1 - F_{\chi_{i_i}}(\lambda_r w_i) \leq \frac{qr}{m}. \quad (5.14)$$

Since $\sigma^{-1} \|\tilde{Y}_{I_i}\|_2 \sim \chi_{i_i}$ for $i \leq m_0$, for such i we have

$$\mathbb{P} \left(w_i^{-1} \|\tilde{Y}_{I_i}\|_2 \geq \sigma \lambda_r \right) = \mathbb{P} \left(\sigma^{-1} \|\tilde{Y}_{I_i}\|_2 \geq \lambda_r w_i \right) = 1 - F_{\chi_{i_i}}(\lambda_r w_i) \leq \frac{qr}{m}. \quad (5.15)$$

Thanks to lemmas 2 and 3, we immediately get

$$\{[\tilde{y}]_I : c_i^* \neq 0 \text{ and } R = r\} \subset \{[\tilde{y}]_I : w_i^{-1} \|\tilde{y}_{I_i}\|_2 > \sigma \lambda_r \text{ and } \tilde{R}^i = r - 1\}, \quad (5.16)$$

which together with (5.15) and (5.16) raises

$$\begin{aligned} \mathbb{P}(c_i^* \neq 0 \text{ and } R = r) &\leq \mathbb{P} \left(w_i^{-1} \|\tilde{Y}_{I_i}\|_2 > \sigma \lambda_r \text{ and } \tilde{R}^i = r - 1 \right) = \\ &\mathbb{P} \left(w_i^{-1} \|\tilde{Y}_{I_i}\|_2 > \sigma \lambda_r \right) \mathbb{P} \left(\tilde{R}^i = r - 1 \right) \leq \\ &\frac{qr}{m} \mathbb{P} \left(\tilde{R}^i = r - 1 \right). \end{aligned} \quad (5.17)$$

Therefore

$$\begin{aligned} \mathbb{E} \left[\frac{V}{\max\{R, 1\}} \right] &= \sum_{r=1}^m \mathbb{E} \left[\frac{V}{r} \mathbf{1}_{\{R=r\}} \right] = \sum_{r=1}^m \frac{1}{r} \mathbb{E} \left[\sum_{i=1}^{m_0} \mathbf{1}_{\{c_i^* \neq 0\}} \mathbf{1}_{\{R=r\}} \right] = \\ &\sum_{r=1}^m \frac{1}{r} \sum_{i=1}^{m_0} \mathbb{P}(c_i^* \neq 0 \text{ and } R = r) \leq \sum_{i=1}^{m_0} \frac{q}{m} \sum_{r=1}^m \mathbb{P}(\tilde{R}^i = r - 1) = \frac{qm_0}{m}, \end{aligned} \quad (5.18)$$

which finishes the proof. ■

In further part of this thesis, we will use the term basic lambdas and use the notation λ^{max} to refer to the sequence of tuning parameters defined in Theorem 13. Figure 3a illustrates the gFDR achieved by gSLOPE under orthogonal case for λ^{max} . In simulation we have fixed 5 groups sizes from the set $\{3, 4, 5, 6, 7\}$, and for each size 200 groups were considered. Consequently, as before we obtained $p = 5000$ and $m = 1000$. Signal sizes were generated such that $\|\beta_{I_i}\|_2 = a\sqrt{l_i}$ for truly relevant groups. Parameter a was chosen to satisfy the condition $\frac{1}{m} \sum_{i=1}^m a\sqrt{l_i} = \frac{1}{m} \sum_{i=1}^m B(m, l_i)$, where $B(m, l)$ is defined in (5.24).

It could be observed, that the selected tuning parameters generate rather conservative settings, i.e. the achieved gFDR is significantly lower than the assumed. This suggests, that penalties (dictated by lambdas) could be slightly decreased, such as the method gets more power and still achieves the gFDR below the assumed level. Returning to the proof of Theorem 13, we can see that the crucial property of defined in the statement of theorem sequence $\{\lambda_i\}_{r=1}^m$ is that $1 - F_{\chi_{i_i}}(\lambda_r w_i) \leq \frac{qr}{m}$ for each i . The possible relaxation of this condition is to assume only that

$$\sum_{i=1}^m \left(1 - F_{w_i^{-1} \chi_{i_i}}(\lambda_r) \right) \leq qr. \quad (5.19)$$

Under equal weights and equal groups sizes scenario, the above conditions are equivalent. In general case, however, the second approach (with the inequality replaced by equality) produces smaller penalties compared to tuning parameters given by Theorem 13. Most often, such a change

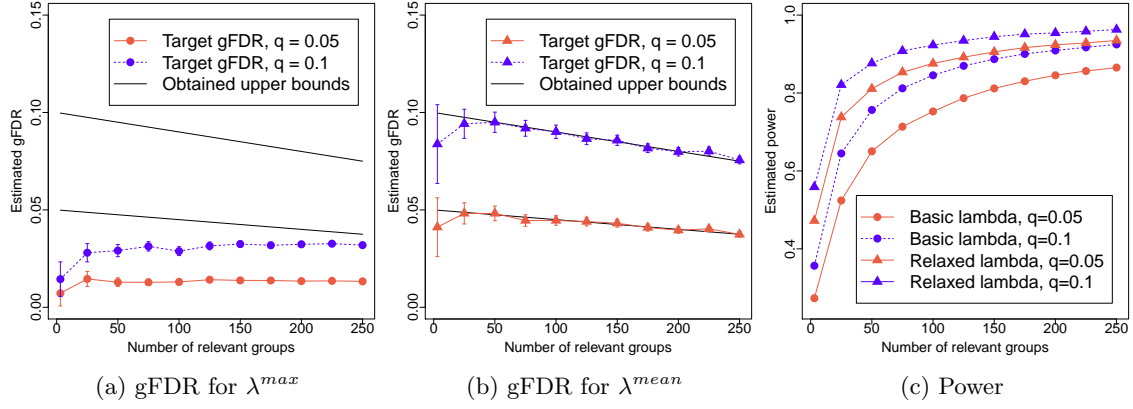


Figure 3: Orthogonal situation with various sizes of groups. For each target gFDR level and true support size, 300 iterations were performed, bars correspond to $\pm 2SE$. Here, black straight lines are given by $q \cdot ((m - k)/m)$, for k being true support size. Weights were defined as $w_i := \sqrt{l_i}$.

results in improving power (and increasing gFDR at the same time). Replacing the inequality in (5.19) by equality yields the following strategy of choosing relaxed λ sequence (denoted by λ^{mean})

$$\lambda_r^{mean} := \bar{F}^{-1} \left(1 - \frac{qr}{m} \right) \quad \text{for} \quad \bar{F}(x) := \frac{1}{m} \sum_{i=1}^m F_{w_i^{-1} \chi_{l_i}}(x), \quad r \in \{1, \dots, m\}, \quad (5.20)$$

where $F_{w_i^{-1} \chi_{l_i}}$ is cumulative distribution function of scaled chi distribution with l_i degrees of freedom and scale $\mathcal{S} = w_i^{-1}$. In Figure 3b we present estimated gFDR, for tuning parameters given by (5.20). The results suggest that with relaxed version of tuning parameters, we can still achieve the gFDR control while essentially improving the power of gSLOPE. Such a strategy could be especially important in situation, when differences between the smallest and the largest quantiles (among distributions $w_i^{-1} \chi_{l_i}$) are relatively large. When this is the case, the gSLOPE with lambdas given by Theorem 13 in orthogonal situation could be considered as too conservative.

5.4 Strength of signals

Consider the case when all groups have the same size, l , and $w > 0$ is used as the universal weight. Assuming the case $X = \mathbf{I}_p$ leads directly to problem (5.12). Here, estimating of the relevant groups by gSLOPE could be summarized as follows: λ decides about the number, R , of groups labeled as relevant; there are chosen R groups which correspond to indices of the R largest values among $w^{-1} \|y_{I_1}\|_2, \dots, w^{-1} \|y_{I_m}\|_2$, being realizations of (possibly) non-central χ distributions with l degrees of freedom and noncentrality parameters given by the entries of $\llbracket \beta \rrbracket_I$. Now, the nonzero $\|\beta_{I_i}\|_2$ could be perceived as strong signal, if with the high probability, the realization of noncentral χ with noncentrality parameter $\|\beta_{I_i}\|_2$ is large comparing to the background composed of the realizations of χ_l distributions (then signal is likely to be identified by gSLOPE; otherwise, the signal could be easily covered by random disturbances and its identification has more in common with good luck, than with the usage of particular method). The important quantity, which could be treated as a breaking point, is the expected value of the maximum of background noise. Groups effects being close to this value, could be perceived as strong under the orthogonal case and medium under the occurrence of weak correlations between groups. The above reasoning applied to the considered case, yields the issue of approximation of the expected value of the maximum of m independent χ_l -distributed variables. Suppose that $\Psi_i \sim \chi_l$ for $i = \{1, \dots, m\}$. From Jensen's

inequality we have

$$\mathbb{E} \left(\max_{i=1, \dots, m} \{\Psi_i\} \right) = \mathbb{E} \left(\sqrt{\max_{i=1, \dots, m} \{\Psi_i^2\}} \right) \leq \sqrt{\mathbb{E} \left(\max_{i=1, \dots, m} \{\Psi_i^2\} \right)},$$

hence we will replace the last problem by the problem of finding the reasonable upper bound on the expected value of the maximum of m independent, χ_l^2 -distributed variables.

Theorem 14. *Let Ψ_1, \dots, Ψ_m be independent variables, $\Psi_i \sim \chi_l^2$ for all i . Then*

$$\mathbb{E} \left(\max_{i=1, \dots, m} \{\Psi_i\} \right) \leq \frac{4 \ln(m)}{1 - m^{-\frac{2}{l}}}. \quad (5.21)$$

Proof. Denote $M_m := \max_{i=1, \dots, m} \{\Psi_i\}$. From the Jensen's inequality applied to e^{tM_m} we have

$$e^{t\mathbb{E}[M_m]} \leq \mathbb{E} [e^{tM_m}] = \mathbb{E} \left[\max_{i=1, \dots, m} e^{t\Psi_i} \right] \leq \sum_{i=1}^m \mathbb{E} [e^{t\Psi_i}]. \quad (5.22)$$

We will consider only $t \in [0, \frac{1}{2})$. Since the moment generating function for χ_l^2 distribution is given by $MGF := m(1-2t)^{-\frac{l}{2}}$, for each i it holds $\mathbb{E} [e^{t\Psi_i}] = (1-2t)^{-\frac{l}{2}}$ and we get $e^{t\mathbb{E}[M_m]} \leq m(1-2t)^{-\frac{l}{2}}$. Applying the natural logarithm to both sides yields

$$\mathbb{E}[M_m] \leq \frac{\ln(m) + \ln \left((1-2t)^{-\frac{l}{2}} \right)}{t}, \quad t \in [0, 1/2). \quad (5.23)$$

Define $t_{m,l} := \frac{1-m^{-\frac{2}{l}}}{2}$. Then for all positive, natural numbers l and m we have $t_{m,l} \in [0, \frac{1}{2})$. Putting $t_{m,l}$ to right side of (5.23) gives inequality (5.21) and finishes the proof. \blacksquare

The above theorem gives us the motivation to use the quantity $\sqrt{4 \ln(m)/(1 - m^{-2/l})}$ as the upper bound on the expected value of maximum over m independent χ_l -distributed variables. The dependence between this function and empirical mean is shown in Figure 4. To estimate expected value of maximum, we performed 10000 iterations for each $m \in \{1, 30, 100, 250, 500, 1000, 2000, 5000\}$. Three degrees of freedom, 5, 20 and 40, were considered. In all simulations, which we have performed to investigate the performance of gSLOPE, we have generated truly relevant groups basing on obtained upper bounds. In particular, in experiments in which sizes of groups as well as weights were identical, we aimed to end with $\mathbb{E} (\|Y_{I_i}\|_2) = \sqrt{4 \ln(m)/(1 - m^{-2/l})}$, for truly relevant group i . Since $\mathbb{E} (\|Y_{I_i}\|_2) \approx \sqrt{\|\beta_{I_i}\|_2^2 + l}$, this yields setting

$$\|\beta_{I_i}\|_2 = B(m, l), \quad \text{for} \quad B(m, l) := \sqrt{4 \ln(m)(1 - m^{-2/l}) - l} \quad (5.24)$$

for groups chosen to be truly relevant.

5.5 The impact of chosen weights

In this subsection we will discuss the influence of chosen weights, $\{w_i\}_{i=1}^m$, on results. Let l_1, \dots, l_m be sizes of groups defined by subsets I_1, \dots, I_m , and assume the orthogonality at groups level, i.e. that it holds $X_{I_i}^\top X_{I_j} = 0$, for $i \neq j$ and that $\sigma = 1$.

Standardization of design matrix makes that condition $X^\top X = \mathbf{I}_p$ is met and we can apply the same reasoning as in subsection 5.3 to get the optimization problem as in (5.11), i.e.

$$c^* = \arg \min_c \frac{1}{2} \|\llbracket \tilde{y} \rrbracket_I - W^{-1}c\|_2^2 + J_\lambda(c), \quad (5.25)$$

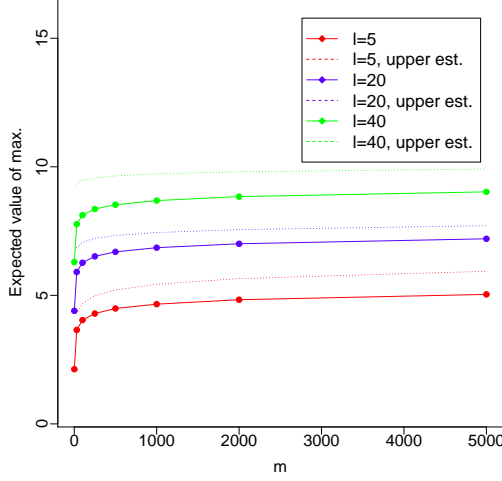


Figure 4: The expected value of the maximum and obtained upper bounds.

where $W_{ii}^{-1} := w_i^{-1}$ and $[\tilde{y}]_I = (\|\tilde{y}_{I_1}\|_2, \dots, \|\tilde{y}_{I_m}\|_2)^\top$. Here, $\tilde{y} = X^\top y$. Suppose now, that c^* has exactly r nonzero coefficients. From Corollary 3, these indices are given by $\{\pi(1), \dots, \pi(r)\}$, where π is permutation which orders $D|\tilde{y}|$. Hence, the order of realizations $\{w_i^{-1}\|\tilde{y}_{I_i}\|_2\}_{i=1}^m$ decides about subset of groups labeled by gSLOPE as relevant. Suppose that groups I_i and I_j are truly relevant, i.e. $\|\beta_{I_i}\|_2 > 0$ and $\|\beta_{I_j}\|_2 > 0$. If we want to achieve the situation in which subset of truly discovered groups is not significantly affected by their sizes, we should choose weights such as $w_i^{-1}\|\tilde{Y}_{I_i}\|_2$ and $w_j^{-1}\|\tilde{Y}_{I_j}\|_2$ are "comparable". One sensible strategy is to look at this issue from the side of expected values. The distributions of $\|\tilde{Y}_{I_i}\|_2$ and $\|\tilde{Y}_{I_j}\|_2$ are noncentral χ distributions, with l_i and l_j degrees of freedom, and the noncentrality parameters equal to $\|\beta_{I_i}\|_2$ and $\|\beta_{I_j}\|_2$, respectively. Now, the expected value of the noncentral χ could be well approximated by square root of the expected value of noncentral χ^2 distribution, which gives

$$\mathbb{E}(w_i^{-1}\|\tilde{y}_{I_i}\|_2) \approx w_i^{-1}\sqrt{\mathbb{E}(\|\tilde{Y}_{I_i}\|_2^2)} = w_i^{-1}\sqrt{l_i + \|\beta_{I_i}\|_2^2}.$$

Therefore, roughly speaking, truly relevant groups I_i and I_j are treated as comparable, when it occurs $l_i/w_i^2 + \|\beta_{I_i}\|_2^2/w_i^2 \approx l_j/w_j^2 + \|\beta_{I_j}\|_2^2/w_j^2$. This gives us the intuition about the behavior of gSLOPE with the choice $w_i = \sqrt{l_i}$ for each i : gSLOPE treats two truly relevant groups as comparable, if groups effect sizes satisfy the condition $\|\beta_{I_i}\|_2/\|\beta_{I_j}\|_2 \approx \sqrt{l_i}/\sqrt{l_j}$. In Figure 5 (results for previously described simulations) we see that when such condition is satisfied, the fractions of groups with different sizes in the selected truly relevant groups (STRG) tend to alignment, when number of all truly relevant groups increases.

The derived condition, $\|\beta_{I_i}\|_2/\|\beta_{I_j}\|_2 = \sqrt{l_i}/\sqrt{l_j}$, could be recast as $\|\beta_{I_i}\|_2^2/l_i = \|\beta_{I_j}\|_2^2/l_j$. This gives a nice interpretation: gSLOPE with the choice $w_i := \sqrt{l_i}$ and under orthogonal situation treats two groups as comparable, when these groups have similar squared effect size of group per coefficient. One possible idealistic situation, when such a property occurs, is when all truly relevant variables have the same impact on the response and were divided into groups containing either all truly relevant or all truly irrelevant regressors.

To investigate the impact of selected weights on the set of discovered groups, we performed simulations with different settings, namely we have put $w_i = 1$ 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 5), when the condition $\|\beta_{I_i}\|_2/\|\beta_{I_j}\|_2 = \sqrt{l_i}/\sqrt{l_j}$ remains unchanged. The results in the

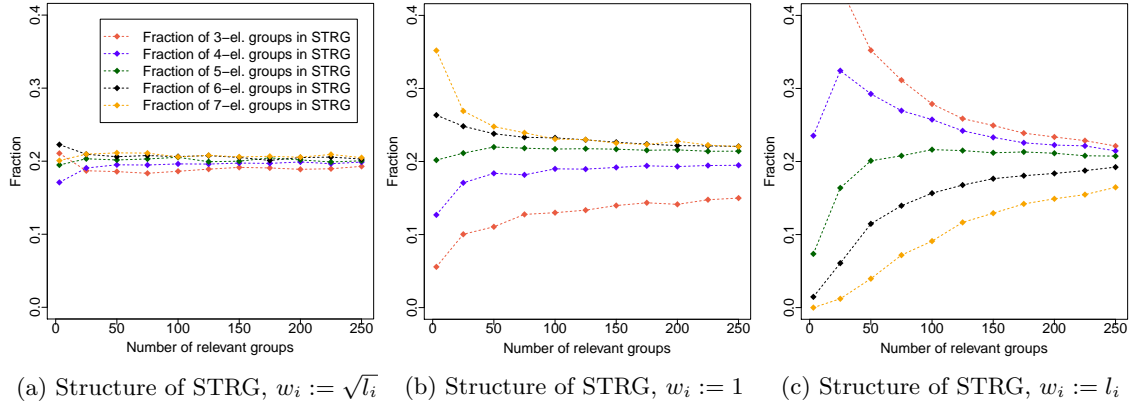


Figure 5: Fraction of each group sizes in STRG. Beyond the weights, this simulation was conducted with the same setting as in experiments summarized in Figure 3 for λ^{mean} . In particular for truly relevant groups i and j it occur $\|\beta_{I_i}\|_2/\|\beta_{I_j}\|_2 = \sqrt{l_i}/\sqrt{l_j}$.

context of gFDR are very similar in all three cases for smoothing parameters adjusted to particular weights. Consequently, we are equipped in the entire family of setting, according to the rule: fix weights to induce the comparability in the domain of group effect strengths for different group sizes, and select lambdas to control gFDR for given weights.

5.6 Near-orthogonal situation

In this section we will deal with the case in which correlations between columns in design matrix are only "almost orthogonal", which could be valid for real-world applications. To simulate such situation we will assume that n by p design matrix is realization of random matrix with entries drawn from the independent normal distributions, $\mathcal{N}(0, \frac{1}{n})$, so as the expected value of $X_i^T X_j$ is equal to 1 for $i = j$, and equal to 0 otherwise. The main objective is to derive the procedure for lambda sequence, which could be applied to achieve gFDR control under assumption that true solution is sparse. In this thesis we will confine ourselves only to case $l_1 = \dots = l_m := l$, $w_1 = \dots = w_m := w$ and number of element in groups is relatively small comparing to number of observations ($l \ll n$). For simplicity in this subsection we will fix $\sigma = 1$. The achieved procedure for generating sequence of lambdas could be then involved in problem (5.1) for any $\sigma > 0$.

In presented in this subsection heuristic, we will use notation $A \approx B$, in order to express that with large probability the differences between corresponding entries of matrices A and B are very small. The procedure for lambdas will be derived after using gSLOPE method directly on X without applying orthogonalization inside groups. Since $X_{I_i}^T X_{I_i} \approx \mathbf{I}_l$, however, both versions give very similar gFDR values.

Assume for simplicity that $\|\beta_{I_1}\|_2 > \dots > \|\beta_{I_s}\|_2 > 0$, $\|\beta_{I_j}\|_2 = 0$ for $j > s$ and solution to gSLOPE, $\tilde{\beta}$, satisfies the same conditions for some λ (this implies in particular, that true signals are relatively strong). Divide I into two families of sets $I^s := \{I_1, \dots, I_s\}$ and $I^c := \{I_{s+1}, \dots, I_m\}$. To derive optimality condition for $\tilde{\beta}$ we will prove following

Theorem 15. *Let $b \in \mathbb{R}^p$ be such that $\|b_{I_1}\|_2 > \dots > \|b_{I_s}\|_2 > 0$, $\|b_{I_j}\|_2 = 0$ for $j > s$ and denote $\lambda^c := (\lambda_{s+1}, \dots, \lambda_m)^T$. If $g \in \partial J_\lambda(w\llbracket b \rrbracket_I)$, then it holds:*

$$\begin{cases} g_{I_i} = w\lambda_i \frac{b_{I_i}}{\|b_{I_i}\|_2}, & i = 1, \dots, s \\ \llbracket g \rrbracket_{I^c} \in C_{w\lambda^c} \end{cases} \quad (5.26)$$

Proof. For $b \in \mathbb{R}^p$ define $J_{\lambda,I}(b) := J_{\lambda}(\llbracket b \rrbracket_I)$ and put $H := \{h \in \mathbb{R}^p : \|(b+h)_{I_1}\|_2 > \dots > \|(b+h)_{I_s}\|_2, \|(b+h)_{I_s}\|_2 > \|h_{I_j}\|_2, j > s\}$. If $g \in \partial J_{\lambda,I}(b)$, then for all $h \in H$, from definition of subgradient

$$\sum_{i=1}^s \lambda_i \|(b+h)_{I_i}\|_2 + \sum_{i=s+1}^m \lambda_i (\llbracket b+h \rrbracket_I)_{(i)} \geq \sum_{i=1}^s \lambda_i \|b_{I_i}\|_2 + \sum_{i=1}^s g_{I_i}^{\top} h_{I_i} + (g^c)^{\top} h^c, \quad (5.27)$$

for $g^c := (g_{I_{s+1}}^{\top}, \dots, g_{I_m}^{\top})^{\top}$ and $h^c := (h_{I_{s+1}}^{\top}, \dots, h_{I_m}^{\top})^{\top}$. Define $\tilde{I} := \{\tilde{I}_1, \dots, \tilde{I}_{m-s}\}$, with set $\tilde{I}_i := \{(i-1) \cdot l + 1, \dots, i \cdot l\}$. Then $\llbracket g^c \rrbracket_{\tilde{I}} = \llbracket g \rrbracket_{I^c}$. Consider first case, when h belongs to the set $H^c := \{h \in H : h_{I_i} \equiv 0, i \leq s\}$. This yields

$$\sum_{i=1}^{m-s} \lambda_{s+i} (\llbracket h^c \rrbracket_{\tilde{I}})_{(i)} \geq (g^c)^{\top} h^c. \quad (5.28)$$

Since $\{h^c : h \in H^c\}$ is open in $\mathbb{R}^{l(m-s)}$ and contains zero, from Corollary 1 we have that $g^c \in \partial J_{\lambda^c, \tilde{I}}(0)$ and the inequality (5.28) is true for any $h^c \in \mathbb{R}^{l(m-s)}$ yielding

$$0 \geq \sup_{h^c} \left\{ (g^c)^{\top} h^c - J_{\lambda^c, \tilde{I}}(h^c) \right\} = J_{\lambda^c, \tilde{I}}^*(g^c). \quad (5.29)$$

The conjugate of $J_{\lambda^c, \tilde{I}}$ is derived in section "Group Ordered Dantzig Selector" (Proposition 17). This result immediately gives condition $\llbracket g^c \rrbracket_{\tilde{I}} \in C_{\lambda^c}$, which is equivalent with $\llbracket g \rrbracket_{I^c} \in C_{\lambda^c}$. To find conditions for g_{I_i} with $i \leq s$, define sets $H_i := \{h \in H : h_{I_j} \equiv 0, j \neq i\}$. For $h \in H_i$, (5.27) reduces to $\lambda_i \|b_{I_i} + h_{I_i}\|_2 \geq \lambda_i \|b_{I_i}\|_2 + g_{I_i}^{\top} h_{I_i}$. Since the set $\{h_{I_i} : h \in H_i\}$ is open in \mathbb{R}^l and contains zero, from Corollary 1 we have $g_{I_i} \in \partial f_i(b_{I_i})$ for $f_i : \mathbb{R}^l \rightarrow \mathbb{R}$, $f_i(x) := \lambda_i \|x\|_2$. Thanks to Proposition 3 (first property), it holds $g_{I_i} = \lambda_i \frac{b_{I_i}}{\|b_{I_i}\|_2}$ for $i \in \{1, \dots, s\}$. The statement follows from the second property from Proposition 3 ■

The third property in Proposition 3 together with Theorems 7 and 10 gives optimality condition for group SLOPE at $\hat{\beta}$:

$$X^{\top}(y - X\hat{\beta}) \in \partial J_{\lambda}(w \llbracket \hat{\beta} \rrbracket_I),$$

which yields

$$\begin{cases} X_{I_i}^{\top}(y - X\hat{\beta}) = w \lambda_i \frac{\hat{\beta}_{I_i}}{\|\hat{\beta}_{I_i}\|_2}, i = 1, \dots, s \\ \llbracket X^{\top}(y - X\hat{\beta}) \rrbracket_{I^c} \in C_{w\lambda^c} \end{cases}. \quad (5.30)$$

After replacing $X_{I_i}^{\top} X_{I_i}$ by \mathbf{I}_l in (5.30), for $i \leq s$ we get $X_{I_i}^{\top}(y - X_{\setminus I_i} \hat{\beta}_{\setminus I_i}) \approx \hat{\beta}_{I_i} \left(1 + \frac{w\lambda_i}{\|\hat{\beta}_{I_i}\|_2}\right)$, where $X_{\setminus I_i}$ is matrix X without columns from I_i and $\hat{\beta}_{\setminus I_i}$ denotes vector $\hat{\beta}$ with removed coefficients indexed by I_i . This means that, for $i = 1, \dots, s$, vector $v_{I_i} := X_{I_i}^{\top}(y - X_{\setminus I_i} \hat{\beta}_{\setminus I_i})$ could be approximated by $\hat{\beta}_{I_i}$ up to a multiplicative constant, which gives $\frac{v_{I_i}}{\|v_{I_i}\|_2} \approx \frac{\hat{\beta}_{I_i}}{\|\hat{\beta}_{I_i}\|_2}$. This yields $\hat{\beta}_{I_i} \approx \left(1 - \frac{w\lambda_i}{\|v_{I_i}\|_2}\right) v_{I_i}$ and consequently $\|\hat{\beta}_{I_i}\|_2 \approx \left| \|v_{I_i}\|_2 - w\lambda_i \right|$. Now from (5.30)

$$\begin{cases} \left| \|v_{I_i}\|_2 - w\lambda_i \right| \approx \|\hat{\beta}_{I_i}\|_2, i = 1, \dots, s \\ \llbracket v \rrbracket_{I^c} \in C_{w\lambda^c} \end{cases}, \quad (5.31)$$

for $v := (v_{I_1}^{\top}, \dots, v_{I_m}^{\top})^{\top}$.

The task now is to select λ_i 's such that condition $\llbracket v \rrbracket_{I^c} \in C_{w\lambda^c}$ regulates the rate of false discoveries. Denote $I_S := \bigcup_{i=1}^s I_i$. Putting $y = X_{I_S} \beta_{I_S} + z$, for truly irrelevant groups ($i > s$)

$$v_{I_i} = X_{I_i}^{\top} X_{I_S} (\beta_{I_S} - \hat{\beta}_{I_S}) + X_{I_i}^{\top} z. \quad (5.32)$$

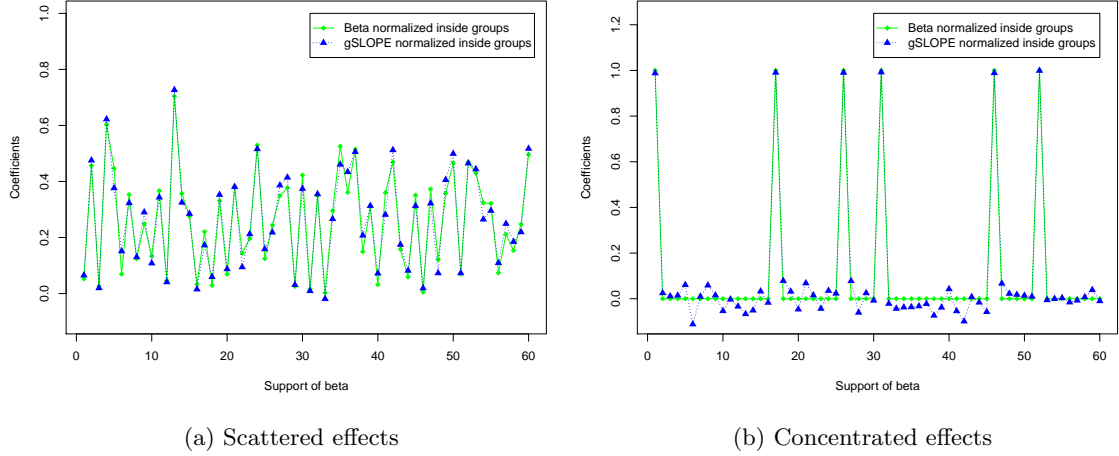


Figure 6: Comparison between relatively strong signals and estimates given by group SLOPE after applying normalization separately for each group. Coefficients with indices greater than 60 were equal to zero (for beta and estimate).

Under orthogonal design this expression reduces only to the term $X_{I_i}^\top z$, and in such situation $\|v_{I_i}\|_2$ has χ distribution with l degrees of freedom which was used in subsection 5.3 to define sequence λ . In considered near-orthogonal situation, the term $X_{I_i}^\top X_{I_S}(\beta_{I_S} - \hat{\beta}_{I_S})$ should be also taken into account. Two following assumptions will be important to derive the appropriate approximation of v_{I_i} distribution:

- the distribution of v_{I_i} could be well approximated by multivariate normal distribution,
- for relatively strong effects it occurs $\frac{\hat{\beta}_{I_i}}{\|\hat{\beta}_{I_i}\|_2} \approx \frac{\beta_{I_i}}{\|\beta_{I_i}\|_2}$ for $i = 1, \dots, s$.

The first assumption is justified when one works with large data scenario. In discussion concerning the second assumption it is important to clarify the effect of penalty imposed on entire groups. The magnitudes of coefficients in $\hat{\beta}_{I_i}$, for truly relevant group i , are generally significantly smaller than in β_{I_i} . This, a so-called shrinking effect, is typical for penalized methods. After normalizing, however, mentioned vectors are often comparable. To demonstrate this effect, we performed simulation with $m = 500$ groups, each containing $l = 10$ elements, and $n = 2000$ observations. First 6 groups were indicated as relevant and for $i = 1, \dots, 6$ we have generated effects such as $\|\beta_{I_i}\|_2 = 25$, which can be perceived as strong signal under such settings. In Figure 6 we present comparison of supports of β and estimate given by group SLOPE for $\lambda_1 = \dots = \lambda_m = 3$, after applying normalization separately for each response related group. We considered scattered effects (6a), as well as whole group signal concentrated at one variable (6b).

From the upper equation in (5.30), we have that $X_{I_S}^\top (X_{I_S} \beta_{I_S} - X_{I_S} \hat{\beta}_{I_S}) + X_{I_S}^\top z \approx w H_{\lambda, \beta}$, for

$$H_{\lambda, \beta} := \left(\lambda_1 \frac{\beta_{I_1}^\top}{\|\beta_{I_1}\|_2}, \dots, \lambda_s \frac{\beta_{I_s}^\top}{\|\beta_{I_s}\|_2} \right)^\top, \quad (5.33)$$

which gives $X_{I_i}^\top X_{I_S}(\beta_{I_S} - \hat{\beta}_{I_S}) \approx X_{I_i}^\top X_{I_S} (X_{I_S}^\top X_{I_S})^{-1} (w H_{\lambda, \beta} - X_{I_S}^\top z)$. Combining the last expression with (5.32) yields

$$v_{I_i} \approx X_{I_i}^\top X_{I_S} (X_{I_S}^\top X_{I_S})^{-1} (w H_{\lambda, \beta} - X_{I_S}^\top z) + X_{I_i}^\top z. \quad (5.34)$$

To determine the parameters of multivariate normal distribution, we will derive the exact values of mean and the covariance matrix of the distribution of right-hand side expression in (5.34) for $i > s$.

Since $I_i \cap I_S = \emptyset$ and entries of X matrix are randomized independently with $\mathcal{N}(0, \frac{1}{n})$ distribution, the expected value of random variable in (5.34) is 0. The second parameter is given by

Theorem 16. *The covariance matrix of $\hat{v}_{I_i} := X_{I_i}^\top X_{I_S} (X_{I_S}^\top X_{I_S})^{-1} (wH_{\lambda, \beta} - X_{I_S}^\top z) + X_{I_i}^\top z$, for $i > s$, is given by the formula*

$$\text{Cov}(\hat{v}_{I_i}) = \left(\frac{n-ls}{n} + w^2 \frac{\|\lambda^S\|_2^2}{n-ls-1} \right) \mathbf{I}_l,$$

where $\lambda^S := (\lambda_1, \dots, \lambda_s)^\top$.

Before proving Theorem 16, we will introduce two lemmas, proofs of which were placed in the Appendix.

Lemma 4. *Suppose that entries of random matrix $X \in M(n, r)$, with $r \leq n$, are independently and identically distributed and have normal distribution with zero mean. Then, there exists expected value of random matrix $A_X = X(X^\top X)^{-1} X^\top$ and it occurs $\mathbb{E}(A_X) = \frac{r}{n} \mathbf{I}_n$.*

Lemma 5. *Suppose that $X \in M(n, r)$, with $r+1 < n$, and entries of X are independent and identically distributed, $X_{ij} \sim \mathcal{N}(0, 1/n)$ for all i and j . Then, there exists expected value of random matrix, $M_{X, \lambda} := B_X H_{\lambda, \beta} H_{\lambda, \beta}^\top B_X^\top$, for $B_X = X(X^\top X)^{-1}$ and $H_{\lambda, \beta}$ defined in (5.33). Moreover, it holds $\mathbb{E}(M_X) = \frac{\|\lambda^S\|_2^2}{n-r-1} \mathbf{I}_n$.*

Proof of Theorem 16. We have $\hat{v}_{I_i} = \xi_{X, z} + \zeta_X$, for $\xi_{X, z} := X_{I_i}^\top (\mathbf{I}_n - A_X) z$, $\zeta_X := wX_{I_i}^\top B_X H_{\lambda, \beta}$, $A_X := X_{I_S} (X_{I_S}^\top X_{I_S})^{-1} X_{I_S}^\top$, $B_X := X_{I_S} (X_{I_S}^\top X_{I_S})^{-1}$. Since $\mathbb{E}(\xi_{X, z} \zeta_X^\top) = 0$ and mean of \hat{v}_{I_i} is equal to 0, it holds $\text{Cov}(\hat{v}_{I_i}) = \text{Cov}(\xi_{X, z}) + \text{Cov}(\zeta_X)$. Now thanks to lemmas 4 and 5

$$\begin{aligned} \text{Cov}(\xi_{X, z}) &= \mathbb{E} \left[X_{I_i}^\top (\mathbf{I}_n - A_X) z z^\top (\mathbf{I}_n - A_X)^\top X_{I_i} \right] = \\ &= \mathbb{E} \left[X_{I_i}^\top (\mathbf{I}_n - A_X) (\mathbf{I}_n - A_X)^\top X_{I_i} \right] = \mathbb{E} \left[X_{I_i}^\top (\mathbf{I}_n - A_X) X_{I_i} \right] = \\ &= \frac{1}{n} (n-ls) \cdot \mathbb{E} \left[X_{I_i}^\top X_{I_i} \right] = \frac{1}{n} (n-ls) \cdot \mathbf{I}_l, \end{aligned} \quad (5.35)$$

$$\begin{aligned} \text{Cov}(\zeta_X) &= w^2 \mathbb{E} \left[X_{I_i}^\top B_X H_{\lambda, \beta} H_{\lambda, \beta}^\top B_X^\top X_{I_i} \right] = w^2 \frac{\|\lambda^S\|_2^2}{n-sl-1} \mathbb{E} \left[X_{I_i}^\top X_{I_i} \right] = \\ &= w^2 \frac{\|\lambda^S\|_2^2}{n-sl-1} \mathbf{I}_l, \end{aligned} \quad (5.36)$$

which finishes the proof. ■

We have shown that for $i > s$ the distribution of $\|v_{I_i}\|_2$ could be approximated by scaled χ distribution with l degrees of freedom and scale parameter $\mathcal{S} = \sqrt{\frac{n-ls}{n} + \frac{w^2 \|\lambda^S\|_2^2}{n-sl-1}}$. Basing on the orthogonal situation, lambdas should be therefore defined as $\lambda_i := \frac{1}{w_i} F_{\chi_l}^{-1} \left(1 - \frac{q \cdot i}{m} \right) = \frac{\mathcal{S}}{w_i} F_{\chi_l}^{-1} \left(1 - \frac{q \cdot i}{m} \right)$. Since s is unknown, we will apply the strategy used in [2]: define λ_1 as in orthogonal case and for $j \geq 2$ define λ_j basing on already generated sequence, according to following procedure.

Procedure 4 Selecting lambdas in near-orthogonal situation: equal groups sizes

input: $q \in (0, 1)$, $w > 0$, $p, n, m, l \in \mathbb{N}$

$\lambda_1 := \frac{1}{w} F_{\chi_l}^{-1} \left(1 - \frac{q}{m} \right)$;

For $i \in \{2, \dots, m\}$:

$\lambda^S := (\lambda_1, \dots, \lambda_{i-1})^\top$;

$\mathcal{S} := \sqrt{\frac{n-l(i-1)}{n} + \frac{w^2 \|\lambda^S\|_2^2}{n-l(i-1)-1}}$;

$\lambda_i^* := \frac{\mathcal{S}}{w_i} F_{\chi_l}^{-1} \left(1 - \frac{q \cdot i}{m} \right)$;

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

To justify the need for correction, when columns in design matrix are only close to be orthogonal, we repeated simulation summarized in Figure 2 with the proviso that this time entries of X were generated from $\mathcal{N}(0, 1/n)$ distribution. In Figure 7 we show gFDR for target levels $q_1 = 0.05$ and $q_2 = 0.1$, when lambdas are chosen basing on (5.13) (basic lambdas, Figure 7a) and when we apply the correction given by Procedure 4 (corrected lambdas, Figure 7b). First 500 coefficients of obtained sequences of smoothing parameters are shown in Figure 7c. Clearly, using basic lambdas,

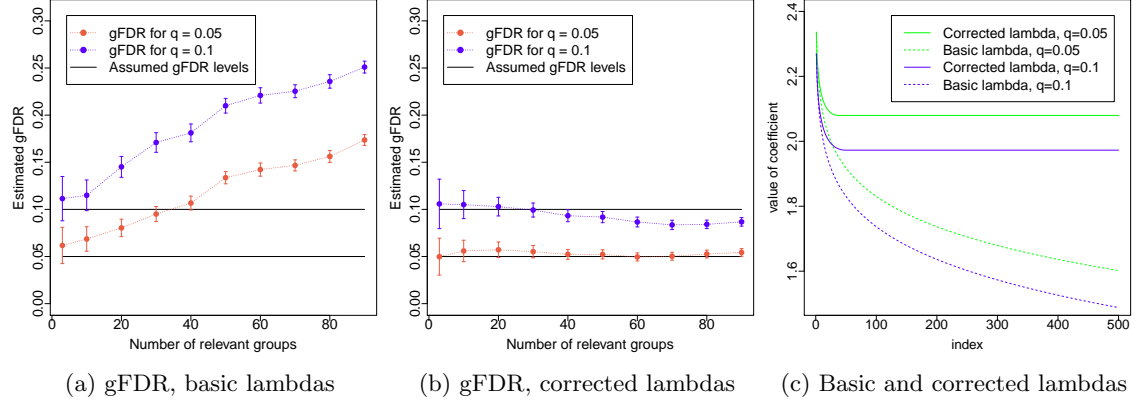


Figure 7: Near-orthogonal situation with equal groups sizes for $l = 5$, $m = 1000$ and $p = n = 5000$. For each target gFDR level and true support size, 200 iterations were performed, bars correspond to $\pm 2SE$. Entries of design matrix were drawn from $\mathcal{N}(0, 1/n)$ distribution and truly relevant signal, i , was generated such as $\|\beta_{I_i}\|_2 = \sqrt{4 \ln(m)/(1 - m^{-2/l}) - l}$.

defined in Theorem 13, results in exceeding of the target gFDR level, even for very sparse $\llbracket \beta \rrbracket_I$. This shows that the correction of smoothing parameters is necessary under occurrence of nonzero correlations between groups.

Consider now the near-orthogonal case with arbitrary groups sizes and sequence of positive weights w_1, \dots, w_m . We can apply an analogous schemes as in orthogonal case. One possible approach is to construct consecutive λ_i as the largest scaled quantiles among all distributions, i.e. as $\max_{j=1, \dots, m} \left\{ \frac{S_j}{w_j} F_{\chi_{l_j}}^{-1} \left(1 - \frac{q^i}{m} \right) \right\}$ for corrections S_j 's adjusted to different l_i values (the conservative strategy). In this thesis, however, we will stick to the more liberal strategy and we will construct relaxed lambdas basing on the concept used earlier in definition of λ^{mean} , (5.20). Therefore we will generate lambdas according to Procedure 5, where the idea is to use the arithmetic mean of scaled

Procedure 5 Selecting lambdas in near-orthogonal situation: arbitrary groups sizes

input: $q \in (0, 1)$, $w_1, \dots, w_m > 0$, $p, n, m, l_1, \dots, l_m \in \mathbb{N}$

$\lambda_1 := \bar{F}^{-1} \left(1 - \frac{q}{m} \right)$, for $\bar{F}(x) := \frac{1}{m} \sum_{i=1}^m F_{w_i^{-1} \chi_{l_i}}(x)$;

for $i \in \{2, \dots, m\}$:

$\lambda^S := (\lambda_1, \dots, \lambda_{i-1})^T$;

$S_j := \sqrt{\frac{n - l_j(i-1)}{n} + \frac{w_j^2 \|\lambda^S\|_2^2}{n - l_j(i-1) - 1}}$, for $j \in \{1, \dots, m\}$;

$\lambda_i^* := \bar{F}_S^{-1} \left(1 - \frac{q^i}{m} \right)$, for $\bar{F}_S(x) := \frac{1}{m} \sum_{j=1}^m \frac{S_j}{w_j} F_{\chi_{l_j}}(x)$;

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

distributions rather than the maximum value which enables to discover more truly relevant groups comparing to the conservative variant. We will test such way of generating lambdas in the next subsection, when we go one step further and assume additionally that the variance of stochastic error is unknown.

5.7 The estimation of the variance of stochastic error, σ^2

Up until this moment, we have used σ in gSLOPE optimization problem, assuming that this parameter is given. 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 , $\hat{\sigma}_{OLS}^2$, given by

$$\hat{\sigma}_{OLS}^2 := (y - X\beta^{OLS})^\top (y - X\beta^{OLS}) / (n - p), \text{ for } \beta^{OLS} := (X^\top X)^{-1} X^\top y. \quad (5.37)$$

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 [2] in the context of SLOPE. Below we present algorithm adjusted to gSLOPE (Procedure 6). The idea standing

Procedure 6 gSLOPE with estimation of σ

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

behind the procedure is simple. 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} , which yields case when numbers of rows exceeds numbers of columns and allows to use (5.37). Algorithm terminates when gSLOPE finds the same subset of relevant variables as in preceding iteration.

For each sparsity level and each from two target gFDR levels, 0.05 and 0.1, 200 iterations were conducted. In each iterations we generated entries of design matrix using $\mathcal{N}(0, \frac{1}{n})$ distribution, then X was standardized and observation were generated according to model (2.7) with $\sigma = 1$. To obtain estimates of relevant groups we have used the iterative version of gSLOPE, with σ estimation (Procedure 6) and lambdas given by Procedure 5. We considered $m = 1000$ groups, with sizes drawn from the binomial distribution, $Bin(1000; 0.008)$, so as the expected value of size was equal to 8 (Figure 8c). We obtained 7917 variables in all groups, n was fixed as 5000. Results were collected in Figure 8.

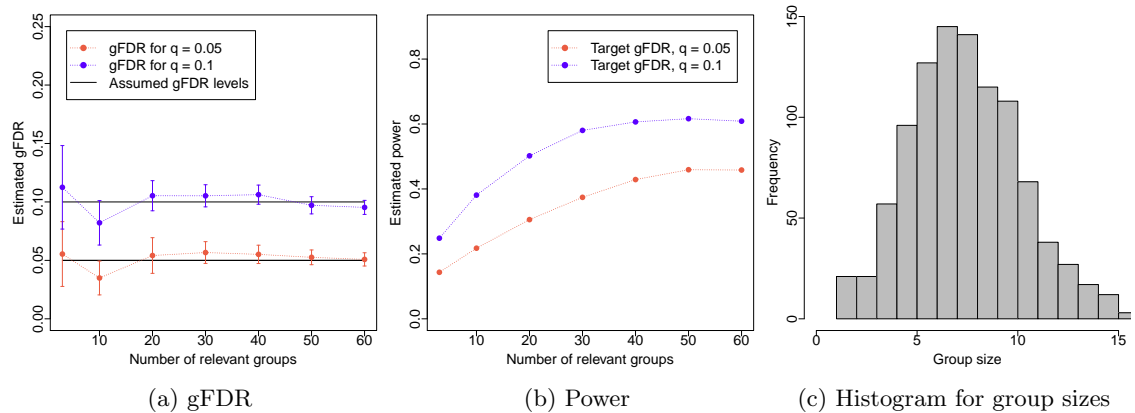


Figure 8: Near-orthogonal situation for various groups sizes with $m = 1000$, $p = 7917$ and $n = 5000$. Bars correspond to ± 2 SE. Entries of design matrix were drawn from $\mathcal{N}(0, 1/n)$ distribution and truly relevant signal, i , was generated such as $\|\beta_{I_i}\|_2 = \frac{1}{m} \sum_{i=1}^m B(m, l_i)$, where $B(m, l)$ is defined in (5.24).

6 Numerical algorithm

In this section we will discuss algorithm for computing the solution to gSLOPE problem (5.1). Our method based on the fast algorithm for evaluation proximity operator (prox) for sorted ℓ_1 norm, which was derived in [2].

Let $I = \{I_1, \dots, I_m\}$ and $\lambda = (\lambda_1, \dots, \lambda_m)^\top$ be vector satisfying $\lambda_1 \geq \dots \geq \lambda_m \geq 0$. Without loss of generality we assume that $\sigma = 1$ in gSLOPE optimization problem. Since considered objective is of form (3.13), we can apply proximal gradient algorithm, provided that norm $J_{\lambda, I, W}$ is prox-capable. To compute the proximal operator for $J_{\lambda, I, W}$ we must be able to minimize function $\frac{1}{2t} \|y - b\|_2^2 + J_{\lambda, I, W}(b)$, for any $y \in \mathbb{R}^p$ and $t > 0$. Multiplying objective by positive number, t , does not change the solution. Such operation leads to new objective function, $\frac{1}{2} \|y - b\|_2^2 + J_{t\lambda, I, W}(b)$. This shows that it is enough to derive fast algorithm solving problem

$$\text{prox}_{J_{\lambda, I, W}}(y) := \arg \min_b \left\{ \frac{1}{2} \|y - b\|_2^2 + J_{\lambda, I, W}(b) \right\}, \quad (6.1)$$

which could be applicable to arbitrary sequence $\lambda_1 \geq \dots \geq \lambda_m \geq 0$.

We will start with situation when W is identity matrix. Simply, then $\text{prox}_J(y)$ is proximal operator for function $J_{\lambda, I}(b) = J_\lambda(\llbracket b \rrbracket_I)$. In such case computing (6.1) could be immediately reduces to finding prox for J_λ norm, since thanks to (5.11) we have

$$\begin{cases} c^* = \arg \min_c \left\{ \frac{1}{2} \|\llbracket y \rrbracket_I - c\|_2^2 + J_\lambda(c) \right\} \\ (\text{prox}_{J_{\lambda, I, W}}(y))_{I_i} = c_i^* (\|y_{I_i}\|_2)^{-1} y_{I_i}, \quad i = 1, \dots, m \end{cases}. \quad (6.2)$$

Consequently, $\text{prox}_J(y)$ could be obtained by applying two steps procedure: find c^* by using fast prox algorithm for J_λ in $\llbracket y \rrbracket_I$, and compute $\text{prox}_{J_{\lambda, I, W}}(y)$ by applying simple calculus to c^* .

Consider now general situation with fixed positive numbers w_1, \dots, w_m and define diagonal matrix M by conditions $M_{I_i, I_i} := w_i^{-1} \mathbf{I}_{I_i}$, for $i = 1, \dots, m$. Then

$$J_{\lambda, I, W}(b) = J_\lambda(W \llbracket b \rrbracket_I) = J_\lambda(\llbracket M^{-1}b \rrbracket_I) = J_{\lambda, I}(M^{-1}b). \quad (6.3)$$

Since M is bijective, we can substitute $\eta := M^{-1}b$ and consider equivalent formulation of (5.1)

$$\begin{cases} \eta^* := \arg \min_\eta \left\{ \frac{1}{2} \|y - XM\eta\|_2^2 + J_{\sigma\lambda, I}(\eta) \right\}, \\ \beta_{gS} = M\eta^* \end{cases}. \quad (6.4)$$

Therefore, after modifying the design matrix, gSLOPE can be always recast as problem with unit weights. Since $J_{\lambda, I}$ is prox-capable, applying proximal gradient method to (6.4) is straightforward. To implement method introduced in this thesis, we have used developed version of an Procedure 3, the accelerated proximal gradient method known as FISTA [32]. In particular, FISTA gives precise procedure for choosing steps sizes, to achieve fast convergence rate. To derive proper stopping criteria, we have considered dual problem to gSLOPE and we have employed the strong duality property. Detailed description was placed in the Appendix.

7 GWAS simulations

DNA is organized into long structures called chromosomes which are almost identical for individuals of a given species. There exist, however, large number of DNA sequences which may occur in different variants (alleles) and single nucleotide polymorphisms, frequently called SNPs, are the most common type of genetic variation. In GWAS one aims in identifying SNPs, which are correlated with an observed trait. Such SNPs show a locations on the genome (quantitative trait locus, QTL)

which host genes that influence a certain quantitative trait. Knowing the number of QTLs related with some phenotypic trait tells us about its genetic architecture, e. g. we can examine if particular trait is controlled by many genes of small effect, or by a few genes of large effect.

People are diploid, i.e. they have two sets of chromosomes. Assuming that two alleles occur, namely a , A for recessive and dominance allele respectively, there exist three possible genotypes: aa , aA and AA which, most commonly, are coded as 0, 1, 2 (the additive coding with respect to recessive allele). Data concerning a given patient are therefore in the form of long sequence of discrete numbers representing genotypes of considered SNPs and are stored in rows of matrix X . Modern technology allows to simultaneously genotype hundreds of thousands of single nucleotide polymorphisms in a single assay, making possible to create a representative samples counting thousands of individuals. After denoting by $\mathbf{1}_n$ the n dimensional column vector of ones, the statistical analysis starts with multiple regression model of form $Y = X\beta + \mu\mathbf{1}_n + z$, with discretely valued matrix X , unknown intercept μ and $z \sim \mathcal{N}(0, \sigma^2 I_n)$. The task is to identify the set of nonzero coefficients of β , which corresponds to selecting response-related SNPs. We have

$$Y = \sum_{i=1}^p X_i \beta_i + \mu \mathbf{1}_n + z = \sum_{i=1}^p (X_i - \bar{X}_i \mathbf{1}_n) \beta_i + \left(\mu + \sum_{i=1}^p \bar{X}_i \beta_i \right) \mathbf{1}_n + z = \tilde{X} \beta + \tilde{\mu} \mathbf{1}_n + z, \quad (7.1)$$

where $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ji}$. Here, \tilde{X} is therefore a matrix obtained after centering columns of X to 0 mean. Left multiplication by $\frac{1}{n} \mathbf{1}_n^T$ yields $\bar{Y} = \tilde{\mu} + \bar{z}$, which gives us the unbiased estimate of $\tilde{\mu}$ parameter, $\hat{\mu} := \bar{Y}$. After replacing $\tilde{\mu}$ by $\hat{\mu}$ in (7.1), we get $\tilde{Y} \approx \tilde{X} \beta + z$, with $\tilde{Y} := Y - \bar{Y} \mathbf{1}_n$. Now,

$$\sum_{i=1}^p \tilde{X}_i \beta_i = \sum_{i=1}^p \tilde{X}_i / \|\tilde{X}_i\|_2 \tilde{\beta}_i, \quad \text{for} \quad \tilde{\beta}_i := \|\tilde{X}_i\|_2 \beta_i, \quad (7.2)$$

hence, since supports of β and $\tilde{\beta}$ are identical, we can normalize columns of \tilde{X} to unit ℓ_2 norm. The above reasoning shows the need for the appropriate mathematical tools applicable in variables selection problem under model (1.1) and justifies the most commonly used assumption about centered and normalized columns in design matrix.

In this section we will apply gSLOPE and SLOPE to genotype data. The goal is to investigate the ability of controlling the fraction of falsely discovered groups with gSLOPE and the ability of controlling the fraction of falsely discovered variables with SLOPE, when real data are considered. We will use the North Finland Birth Cohort (NFBC) dataset, described in detail in [36]. In our analysis we will restrict only to SNPs from chromosome 1. Simulations were performed as follows.

- Columns in X were centered and normalized, SNPs were clustered into m groups basing on correlations between columns.
- In each iteration for assumed sparsity level, k , we have randomized k separated (i.e. the minimum range between groups was set as 60 indices) groups, which were treated as truly relevant.
- We have put all group effect into only one variable in group, so as the number of all truly relevant groups was equal to the number of all truly relevant variables, i.e. we have considered situation with the same sparsity at variables and groups levels.
- We have put $\beta_i := \frac{1}{m} \sum_{i=1}^m B(m, l_i)$, for truly relevant variable i and $\beta_i := 0$, otherwise; model (1.1) with $\sigma = 1$ was used to generate observations, y .
- The estimates of gFDR for gSLOPE and FDR for SLOPE were obtained after using both methods to the same y in each iteration. SLOPE was applied directly to design matrix, while in gSLOPE case we have orthogonalized columns inside groups first (i.e. standardization described in subsection 5.2 was applied).

To define groups in considered set of SNPs, the hierarchical cluster algorithm (HCA) [37], [38] was employed. HCA is a technique which builds a hierarchy of clusters for p variables, basing on the p by p dimensional distance matrix. This matrix was constructed according to correlations between variables, i.e. the distance between centered and normalized columns i and j was defined as $d(X_i, X_j) := 1 - |X_j^\top X_i|$, so as the maximal distance, 1, is taken for orthogonal predictors, and minimal distance, 0, for identical columns.

The HCA starts by assigning each item to a cluster. Thus, the first step produces p clusters, each containing just one item. In next steps, the distances between current clusters are calculated, i.e. the values $\delta(i, j) = \frac{1}{|C_1| \cdot |C_2|} \sum_{X_i \in C_1} \sum_{X_j \in C_2} d(X_i, X_j)$ for each i and j . The pair of clusters with the smallest value $\delta(i, j)$ is merged into a single cluster and the procedure continues until some stop criterion (e.g. predefined number of output clumps) is satisfied.

The correlation structure of starting data is shown in Figure 9a. This set of SNPs was divided into 3000 groups and then very large clusters (more than 15 elements) were removed. As a result, we have obtained the design matrix, X , with $n = 5402$ rows and $p = 12677$ columns, and groups I_1, \dots, I_m , for $m = 2522$. Columns were permuted, so as the groups were arranged one after another, i.e. $I_1 = \{1, \dots, l_1\}$, $I_2 = \{l_1 + 1, \dots, l_2\}$ and so on. The correlations inside and between groups are shown in Figure 9b, while the histogram showing the proportion of each group size in resulting division is shown in Figure 10c. Mean cluster size was equal to 5.03.

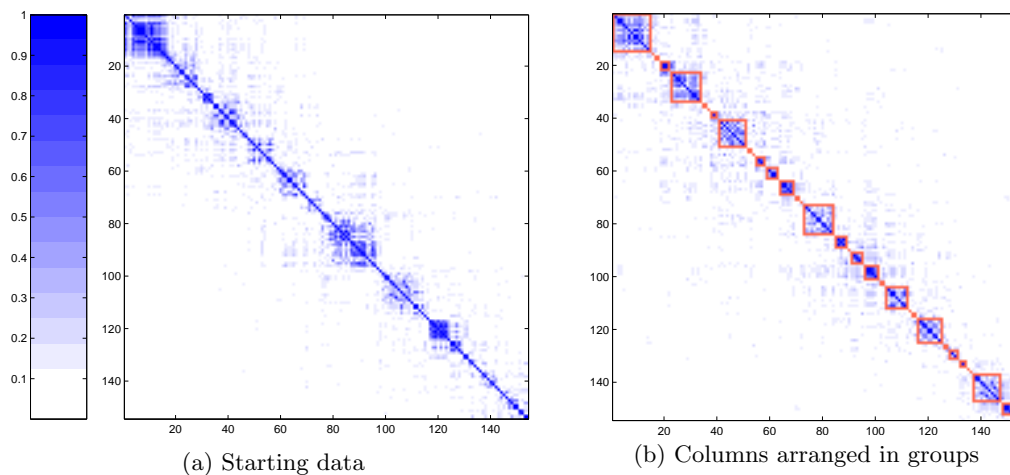


Figure 9: Heat maps of correlation matrices (absolute values) showing the structure in design matrix. The HCA method was used to define clusters of strongly correlated variables. Red frames show the division of columns into groups.

To generate observations, the stochastic errors were drawn from standard normal distribution. When tested methods were applied, however, we have assumed that σ is unknown parameter, which needs to be estimated. To select truly relevant variables, we have used the iterative version of SLOPE introduced in [2]. The selection of relevant groups was made by gSLOPE (Procedure 6), with lambdas given by Procedure 5. Estimated FDR for SLOPE is shown in Figure 10a, while obtained gFDR for gSLOPE method is presented in Figure 10b.

Despite the fact that between closely located groups there still occur correlations suggesting statistical relationships, using gSLOPE with heuristic derived for near-orthogonal case, with $X_{ij} \sim \mathcal{N}(0, 1/n)$, gives gFDR only slightly above the assumed level. Estimated FDR for SLOPE is significantly higher than target level q , which follows from the SLOPE tendency to select strongly correlated predictors together with (or instead of) the truly relevant. The obtained results show the potential of the gSLOPE in GWAS, when fraction of falsely identified groups in all discoveries is treated as the criterion for selection. We want to emphasise, however, that achieved results sub-

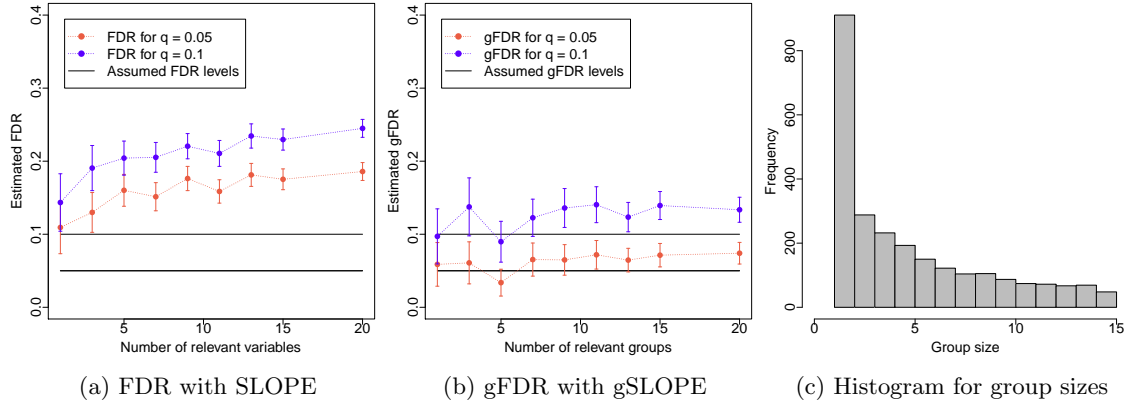


Figure 10: The usage of gSLOPE and SLOPE in genetic application. In simulations we have considered 12677 SNPs, divided by HCA into 2522 groups, and 5402 subjects. The iterative SLOPE and gSLOPE with σ estimation were applied to select relevant predictors or groups. Two target FDR/gFDR levels were tested, $q = 0.05$ and $q = 0.1$.

stantially depend on chosen clustering algorithm, and HCA is only one possible choice. Testing for the most appropriate way of dividing SNPs into groups seems to be very important point in further investigation on the usage of gSLOPE under genetic applications.

8 Group Ordered Dantzig Selector

Consider unconstrained optimization problem of form

$$\underset{b}{\text{minimize}} \quad f(b) = g(b) + h(b), \quad (8.1)$$

where $g : \mathbb{R}^p \rightarrow \mathbb{R}$, $h : \mathbb{R}^p \rightarrow \mathbb{R}$ are convex functions, g is differentiable and h is some sparsity inducing norm. Since f is subdifferentiable, from KKT conditions we have $-\nabla g(b^*) \in \partial h(b^*)$, for b^* being solution to (8.1).

Proposition 16. *Let h^D be dual norm to h and suppose that $v \in \partial h(b)$ for $b \in \mathbb{R}^p$. Then it occurs*

$$h^D(v) \leq 1. \quad (8.2)$$

Proof. We have $h(b) = \sup_{\alpha} \{\alpha^T b : h^D(\alpha) \leq 1\}$. Therefore, from Theorem 8,

$$\begin{aligned} \partial h(b) &= \text{conv} \bigcup \{\partial(\alpha^T b) : \alpha^T b = h(b), h^D(\alpha) \leq 1\} = \\ &= \text{conv} \bigcup \{\alpha : \alpha^T b = h(b), h^D(\alpha) \leq 1\} \subset \{\alpha : h^D(\alpha) \leq 1\}, \end{aligned}$$

where inclusion follows from the convexity of set $\{\alpha : h^D(\alpha) \leq 1\}$. ■

It is easy to see, that LASSO optimization problem (2.1) is of form (8.1), with $g(b) = \frac{1}{2} \|y - Xb\|_2^2$ and $h(b) = \sigma \lambda_L \|b\|_1$. Since it holds $h^D(b) = \sigma^{-1} \lambda_L^{-1} \|b\|_\infty$, condition (8.2) with $v = -\nabla g(b^*)$ gives $\|X^T(y - Xb^*)\|_\infty \leq \sigma \lambda_{DS}$, which coincides with the feasible set of DS estimate (2.2) with $\lambda_{DS} = \lambda_L$. This shows the manner how the DS approach could be generalized for any problem of form (8.1): use sparsity inducing norm h as objective in constrained convex optimization problem with feasible set given by (8.2).

We will apply the DS approach in the gSLOPE context. The first step is to find the dual norm to $J_{\lambda,I,W}$. We have $J_{\lambda,I,W}^D(x) = \max_b \{x^\top b : J_{\lambda,I,W}(b) \leq 1\} = \max_b \{x^\top b : J_\lambda(W\llbracket b \rrbracket_I) \leq 1\} = \max_{b,c} \{x^\top b : J_\lambda(c) \leq 1, c = W\llbracket b \rrbracket_I\} = \max_c \{x^\top b^c : J_\lambda(c) \leq 1, c \succeq 0\}$, where b^c is defined as

$$b^c := \arg \max_b \{x^\top b : c = W\llbracket b \rrbracket_I\}. \quad (8.3)$$

This problem is separable and for each i we have $b_{I_i}^c = \arg \max \{x_{I_i}^\top b_{I_i} : \frac{1}{2}c_i^2 = \frac{1}{2}w_i^2\|b_{I_i}\|_2^2\}$. The principle of Lagrange multiplier (with the Lagrange multiplier μ_i) gives conditions

$$x_{I_i} = \mu_i w_i^2 b_{I_i}^c, \quad c_i^2 = w_i^2 \|b_{I_i}^c\|_2^2. \quad (8.4)$$

Multiplying both sides of the stationary condition by $(b_{I_i}^c)^\top$ gives $x_{I_i}^\top b_{I_i}^c = \mu_i c_i^2$. On the other hand, multiplying the same equation by $c_i^2 x_{I_i}^\top$, yields $c_i^2 \|x_{I_i}\|_2^2 = w_i^2 \mu_i c_i^2 x_{I_i}^\top b_{I_i}^c = w_i^2 (x_{I_i}^\top b_{I_i}^c)^2$. Therefore, it holds $x_{I_i}^\top b_{I_i}^c = c_i w_i^{-1} \|x_{I_i}\|_2$ or $x_{I_i}^\top b_{I_i}^c = -c_i w_i^{-1} \|x_{I_i}\|_2$ and, clearly, the first case corresponds to global maximum. Consequently,

$$\begin{aligned} J_{\lambda,I,W}^D(x) &= \max_c \{(W^{-1}\llbracket x \rrbracket_I)^\top c : J_\lambda(c) \leq 1, c \succeq 0\} = \\ &= \max_c \{(W^{-1}\llbracket x \rrbracket_I)^\top c : J_\lambda(c) \leq 1\} = J_\lambda^D(W^{-1}\llbracket x \rrbracket_I) \end{aligned} \quad (8.5)$$

and from Proposition 9, $J_{\lambda,I,W}^D(x) = \max \left\{ \lambda_1^{-1} |W^{-1}\llbracket x \rrbracket_I|_{(1)}, \dots, \left(\sum_{i=1}^p \lambda_i \right)^{-1} \sum_{i=1}^p |W^{-1}\llbracket x \rrbracket_I|_{(i)} \right\}$. Theorem 4 gives important result about the conjugate of group sorted ℓ_1 norm, which we shall use more than once.

Proposition 17. *The conjugate function for $J_{\lambda,I,W}$ is the function $J_{\lambda,I,W}^*$ defined as*

$$J_{\lambda,I,W}^*(x) = \begin{cases} 0, & W^{-1}\llbracket x \rrbracket_I \in C_\lambda \\ \infty, & \text{otherwise} \end{cases}. \quad (8.6)$$

We are now in a position to introduce the generalization of DS in the context of gSLOPE. For given sequence of nonincreasing, nonnegative starting parameters $\lambda_1, \dots, \lambda_m$, observations y , groups $I = \{I_1, \dots, I_m\}$ and given sequence of positive weights w_1, \dots, w_m , group Ordered Dantzig Selector (gODS) is defined as the solution to

$$\beta^{\text{gODS}} := \arg \min_b J_{\lambda,I,W}(b) \quad \text{subject to} \quad W^{-1}\llbracket X^\top(y - Xb) \rrbracket_I \in C_{\sigma\lambda}, \quad (8.7)$$

where W is diagonal matrix defined by equations $W_{i,i} := w_i$, for $i = 1, \dots, m$. For β^{gODS} being the solution to above optimization problem, we define the estimate of $\llbracket \beta \rrbracket_I$ support by the indices corresponding to nonzeros of $\llbracket \beta^{\text{gODS}} \rrbracket_I$. Under single-variable-in-groups scenario, with $w_i = 1$ for all i , we will refer to solution of (8.7) as Ordered Dantzig Selector (ODS). ODS could be treated as alternative method to SLOPE and used in model selection problem. In this thesis we will investigate the issue of connections between ODS and SLOPE under orthogonal design. Our finding provides a basis for further studies on the relationship between gODS and gSLOPE. Our main result is

Theorem 17. *Suppose that λ satisfies $\lambda_1 > \dots > \lambda_p > 0$ and $X^\top X = \mathbf{I}_p$ for $X \in M(n, p)$. Then solutions to problems*

$$\arg \min_b J_\lambda(b) \quad \text{subject to} \quad X^\top(y - Xb) \in C_{\sigma\lambda} \quad (\text{ODS})$$

and

$$\arg \min_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + J_{\sigma\lambda}(b) \right\} \quad (\text{SLOPE})$$

are unique and identical. In particular, this result provides the FDR control on arbitrary level $q \in (0, 1)$ for ODS, with λ^{BH} sequence, defined as in Theorem 1, i.e. it holds $FDR \leq \frac{qp_0}{p}$, where p_0 is number of zero coefficients in β .

Before proving the above theorem, we will recall results concerning (SLOPE) and derive some properties useful in analysis of problem (ODS). Since $\frac{1}{2}\|y - Xb\|_2^2 = \frac{1}{2}\|\tilde{y} - b\|_2^2 + \text{const}$ and $X^\top(y - Xb) = \tilde{y} - b$, with $\tilde{y} := X^\top y$, without loss of generality we will consider the case $X = \mathbf{I}_p$ and $\sigma = 1$. Let I_1 be some subset of $\{1, \dots, p\}$ and $\overline{b_{I_1}}$ denote the arithmetic mean of subvector b_{I_1} , for any $b \in \mathbb{R}^p$. Moreover we will use notation b_+ for nonnegative vector given by conditions $(b_+)_i := \max\{b_i, 0\}$. It was shown in [2] that the unique solution to (SLOPE) (uniqueness follows from strict convexity of objective function) is output of the following procedure, which terminates in at most n steps.

Procedure 7 Solution to SLOPE in the orthogonal case [2]

Input: Nonnegative and nonincreasing sequences y and λ .

while $y - \lambda$ is not nonincreasing **do**

Identify strictly increasing subsequences, i.e. segments $I_i := \{j, \dots, l\}$ such that

$$y_j - \lambda_j < \dots < y_l - \lambda_l.$$

For each $k \in I_i$ replace the values of y and λ by their average value over such segments

$$y_k \leftarrow \overline{y_{I_i}}, \quad \lambda_k \leftarrow \overline{\lambda_{I_i}}$$

end while

output: $\beta^S = (y - \lambda)_+$.

Authors of [2] have showed that (SLOPE) could be recast as quadratic program of form

$$\begin{aligned} & \text{minimize} && \frac{1}{2}\|y - b\|_2^2 + \lambda^\top b \\ & \text{subject to} && b_1 \geq \dots \geq b_p \geq 0. \end{aligned} \tag{8.8}$$

The proof of Theorem 17 is constructed as follows. After showing that all solutions to (ODS) satisfy certain conditions, we will reduce (ODS) to linear program and we will show that solution to achieved problem is unique. Then, we will focus on some special perturbation of linear problem (which enables to omit the assumption about strict monotonicity of λ in further part of proof), with the same solution as original problem, and we will show that this solution is given by Procedure 7, which implies the oneness with solution to SLOPE.

Including the assumption $X = \mathbf{I}_p$ yields

$$\arg \min_b J_\lambda(b) \quad \text{subject to} \quad y - b \in C_\lambda. \tag{8.9}$$

It is easy to see, that if $y - b \in C_\lambda$, then also $P_\pi(y - b) \in C_\lambda$ and $S(y - b) \in C_\lambda$, for any permutation π and diagonal matrix S with $|S_{ii}| = 1$, $i = 1, \dots, p$.

Proposition 18. *Suppose that b^* is solution to (8.9), π is arbitrary permutation of $\{1, \dots, p\}$ and S is diagonal matrix such as $|S_{ii}| = 1$ for all i . Then*

i) $b_\pi^ := P_\pi b^*$ is solution to $\arg \min_b J_\lambda(b)$ s. t. $P_\pi y - b \in C_\lambda$,*

ii) $b_S^ := S b^*$ is solution to $\arg \min_b J_\lambda(b)$ s. t. $S y - b \in C_\lambda$.*

Proof. Suppose that there exists $b_0 \in \mathbb{R}^p$ such as $J_\lambda(b_0) < J_\lambda(b_\pi^*)$ and $P_\pi y - b_0 \in C_\lambda$. Then we have $y - P_\pi^\top b_0 \in C_\lambda$ and $J_\lambda(P_\pi^\top b_0) < J_\lambda(b^*)$, which contradicts the optimality of b^* . Analogously, the second part of proposition could be proved. ■

Thanks to Proposition 18, without loss of generality we will assume that $y_1 \geq \dots \geq y_p \geq 0$. Indeed, if for arbitrary $y \in \mathbb{R}^p$, the solution, $b^{|y|(\cdot)}$, to (8.9) for ordered magnitudes of observations is known, the original solution could be immediately recovered as $b^* = SP_\pi b^{|y|(\cdot)}$, where P_π and S are selected to satisfy $y = SP_\pi |y|(\cdot)$. This coincides with the analogous property for SLOPE.

Proposition 19. *Suppose that $\lambda_1 \geq \dots \geq \lambda_p \geq 0$. For arbitrary $x \in \mathbb{R}^p$, $\varepsilon > 0$ and $l, j \in \{1, \dots, p\}$ define:*

$$(x_\varepsilon)_i = \begin{cases} x_l + \varepsilon, & i = l \\ x_j - \varepsilon, & i = j \\ x_i, & \text{otherwise} \end{cases}, \quad (\tilde{x}_\varepsilon)_i = \begin{cases} x_j - \varepsilon, & i = j \\ x_i, & \text{otherwise} \end{cases}. \quad (8.10)$$

Then:

i) if $x_j > x_l \geq 0$, then for $\varepsilon \in (0, (x_j - x_l)/2]$ it holds $J_\lambda^D(x_\varepsilon) \leq J_\lambda^D(x)$,

ii) if $x_j > 0$, then for $\varepsilon \in (0, x_j]$ it holds $J_\lambda^D(\tilde{x}_\varepsilon) \leq J_\lambda^D(x)$.

Proof. It is easy to observe that for any $x \in \mathbb{R}^p$, the dual to sorted ℓ_1 norm could be represented as $J_\lambda^D(x) = \max \{J_{\lambda^k}(x), k \leq p\}$, for $\lambda_i^k := \begin{cases} (\sum_{j=1}^k \lambda_j)^{-1}, & i \leq k \\ 0, & \text{otherwise} \end{cases}$. The claim is therefore the straightforward consequence of Propositions 7 and 8. \blacksquare

Proposition 20. *If $|x| \leq |\tilde{x}|$, then $J_\lambda^D(x) \leq J_\lambda^D(\tilde{x})$.*

Proof. Follows simply from Corollary 2, analogously as in proof of previous proposition. \blacksquare

Proposition 21. *Assume that $\lambda \succ 0$, $y_1 \geq \dots \geq y_p \geq 0$ and let b^* be solution to (8.9). Then, for any $j \in \{1, \dots, p\}$ we have $0 \leq b_j^* \leq y_j$.*

Proof. Suppose first, that for some j it occurs $b_j^* < 0$ and define $(b_\varepsilon)_i := \begin{cases} |b_j^*| - \varepsilon, & i = j \\ |b_i^*|, & \text{otherwise} \end{cases}$. Fix $\varepsilon = |b_j^*|$. Then $|y - b_\varepsilon| \leq |y - b^*|$, hence Proposition 20 yields $J_\lambda^D(y - b_\varepsilon) \leq J_\lambda^D(y - b^*) \leq 1$ and b_ε is feasible. Moreover, Proposition 7 gives $J_\lambda(b_\varepsilon) < J_\lambda(|b^*|) = J_\lambda(b^*)$, which leads to contradiction.

Suppose now, that $b_j^* > y_j$. This gives that $b_j^* > 0$. Define $(b_\varepsilon)_i := \begin{cases} b_j^* - \varepsilon, & i = j \\ b_i^*, & \text{otherwise} \end{cases}$ and fix $\varepsilon := b_j^* - y_j$. As before b_ε is feasible. Using again Proposition 7, we get $J_\lambda(b_\varepsilon) < J_\lambda(b^*)$ which contradicts the optimality of b^* . \blacksquare

Proposition 22. *Assume that $\lambda_1 > \dots > \lambda_p > 0$, $y_1 \geq \dots \geq y_p \geq 0$ and let b^* be solution to (8.9). Then it occurs $b_1^* \geq \dots \geq b_p^* \geq 0$ and $y_1 - b_1^* \geq \dots \geq y_p - b_p^* \geq 0$.*

Proof. From previous propositions we have that $b^* \succeq 0$ and $y - b^* \succeq 0$. We will show that for $1 \leq j < l \leq p$ it holds $b_j^* \geq b_l^*$ and $y_j - b_j^* \geq y_l - b_l^*$. Suppose first, that $b_j^* < b_l^*$ for some $j < l$ and denote $x := y - b^*$. Since $y_j \geq y_l$, we have $x_j > x_l$. Define

$$(b_\varepsilon)_i = \begin{cases} b_j^* + \varepsilon, & i = j \\ b_l^* - \varepsilon, & i = l \\ b_i^*, & \text{otherwise} \end{cases}, \quad (x_\varepsilon)_i = \begin{cases} x_j - \varepsilon, & i = j \\ x_l + \varepsilon, & i = l \\ x_i, & \text{otherwise} \end{cases}. \quad (8.11)$$

Then $x_\varepsilon = y - b_\varepsilon$. From Propositions 8 and 19, there exist $t_1, t_2 > 0$ such that $J_\lambda(b_\varepsilon) < J_\lambda(b^*)$ for $\varepsilon \in (0, t_1)$ and $J_\lambda^D(x_\varepsilon) \leq J_\lambda^D(x)$ for $\varepsilon \in (0, t_2)$. Define $t := \min\{t_1, t_2\}$ and fix $\varepsilon \in (0, t)$. Then, we have $J_\lambda^D(y - b_\varepsilon) = J_\lambda^D(x_\varepsilon) \leq J_\lambda^D(x) \leq 1$, hence b_ε is feasible with smaller value of objective.

Suppose now, that $y_j - b_j^* < y_l - b_l^*$ for $j < l$. This gives $x_j < x_l$ and $b_j^* > b_l^*$. The feasible vector, b_ε , with smaller value of objective, could be now constructed in an analogous manner, again yielding the contradiction with optimality of b^* . \blacksquare

Proposition 22 states that including new inequality constraints, i.e. $b_1 \geq \dots \geq b_p \geq 0$ and $y_1 - b_1 \geq \dots \geq y_p - b_p \geq 0$, to problem (8.9) does not change the set of solutions. These additional restraints simplify objective function and as a result the task takes form of minimizing linear function $\lambda^\top b$. Moreover condition $y - b \in C_\lambda$, could be now represented by p affine constraints of form $\sum_{i=1}^k (y_i - \lambda_i) \leq \sum_{i=1}^k b_i$. Summarizing, (8.9) could be recast as linear program. We will now show that after such transformation, one can omit the conditions $y_1 - b_1 \geq \dots \geq y_p - b_p \geq 0$, yielding equivalent formulation

$$\begin{aligned} & \underset{b}{\text{minimize}} && \lambda^\top b \\ & \text{s.t.} && \begin{cases} \sum_{i=1}^k (y_i - \lambda_i) \leq \sum_{i=1}^k b_i, & k = 1, \dots, p \\ b_1 \geq \dots \geq b_p \geq 0 \end{cases} \end{aligned} \quad (8.12)$$

Proposition 23. *Let b^* be solution to (8.12), for $\lambda_1 > \dots > \lambda_p > 0$ and $y_1 \geq \dots \geq y_p \geq 0$. Then*

$$\begin{aligned} & i) \quad \sum_{i=1}^j (y_i - \lambda_i) = \sum_{i=1}^j b_i^* \quad \text{or} \quad b_j^* = b_{j+1}^*, \\ & ii) \quad y_j - b_j^* \geq y_{j+1} - b_{j+1}^*, \end{aligned} \quad (8.13)$$

for all $j \in \{1, \dots, p\}$, with the convention that $b_{p+1}^* := 0$ and $y_{p+1} := 0$.

Proof. Fix $j \in \{1, \dots, p\}$ and suppose that $b \in \mathbb{R}^p$ is feasible vector for problem (8.12) such as $\sum_{i=1}^j (y_i - \lambda_i) < \sum_{i=1}^j b_i$ and $b_j > b_{j+1}$, with the convention that $b_{p+1} = 0$. There exists $\varepsilon > 0$, such as

$$\sum_{i=1}^j (y_i - \lambda_i) < \left(\sum_{i=1}^j b_i \right) - \varepsilon \quad \text{and} \quad b_j - \varepsilon > b_{j+1} + \varepsilon. \quad (8.14)$$

Define $b_\varepsilon \in \mathbb{R}^p$ by putting $(b_\varepsilon)_j := b_j - \varepsilon$, $(b_\varepsilon)_{j+1} := b_{j+1} + \varepsilon$ and $(b_\varepsilon)_i := b_i$ for $i \notin \{j, j+1\}$. Thanks to (8.14), b_ε is feasible (note that $\sum_{i=1}^k b_i = \sum_{i=1}^k (b_\varepsilon)_i$, for $k \neq j$). Now, with convention $\lambda_{p+1} := 0$, it holds $\lambda^\top b - \lambda^\top b_\varepsilon = \varepsilon(\lambda_j - \lambda_{j+1}) > 0$, which shows that b is not optimal.

To prove ii), let b be any feasible point, such that $y_j - b_j < y_{j+1} - b_{j+1}$ for some $j \in \{1, \dots, p\}$. Considering the case $j = 1$, from feasibility of b we get $y_1 - b_1 < \frac{(y_1 - b_1) + (y_2 - b_2)}{2} \leq \frac{\lambda_1 + \lambda_2}{2} < \lambda_1$. For $j \in \{2, \dots, p\}$ we have $\sum_{i=1}^{j-1} (y_i - b_i) \leq \sum_{i=1}^{j-1} \lambda_i$, $\sum_{i=1}^{j+1} (y_i - b_i) \leq \sum_{i=1}^{j+1} \lambda_i$ (with $\lambda_{p+1} := 0$). Adding both sides of these inequalities and dividing by 2 yields

$$\sum_{i=1}^{j-1} (y_i - b_i) + \frac{(y_j - b_j) + (y_{j+1} - b_{j+1})}{2} \leq \sum_{i=1}^{j-1} \lambda_i + \frac{\lambda_j + \lambda_{j+1}}{2} < \sum_{i=1}^j \lambda_i. \quad (8.15)$$

Therefore thanks to assumption $y_j - b_j < y_{j+1} - b_{j+1}$:

$$\sum_{i=1}^j (y_i - b_i) < \sum_{i=1}^{j-1} (y_i - b_i) + \frac{(y_j - b_j) + (y_{j+1} - b_{j+1})}{2} < \sum_{i=1}^j \lambda_i. \quad (8.16)$$

Summarizing, we always have $\sum_{i=1}^j (y_i - \lambda_i) < \sum_{i=1}^j b_i$. Moreover, $y_j \geq y_{j+1}$ and $y_j - b_j < y_{j+1} - b_{j+1}$ give that $b_j > b_{j+1}$. Therefore, from i), vector b can not be optimal. \blacksquare

We will now show that problem (8.12) has unique solution. For $k \in \mathbb{N}$, define k by k , upper and lower triangular matrices S_k and V_k as

$$(S_k)_{j,l} = \begin{cases} 1, & j \leq l \\ 0, & \text{otherwise} \end{cases}, \quad (V_k)_{j,l} = \begin{cases} 1, & l = j \\ -1, & j = l + 1 \\ 0, & \text{otherwise} \end{cases}. \quad (8.17)$$

It could be easily verified, that $S^{-1} = V^\top$. We are now ready to prove

Lemma 6. *Solution to (8.12) is unique, for $\lambda_1 > \dots > \lambda_p > 0$.*

Proof. Denote the columns of matrices S_p and V_p by s_1, \dots, s_p and v_1, \dots, v_p , respectively. From the condition $S_p V_p^\top = \mathbf{I}_p$, we get that s_i is orthogonal to v_j for $i \neq j$. Hence, since matrices S_p and V_p are nonsingular, the matrix $[(S_p)_{I_1} | (V_p)_{I_2}]$ is nonsingular as well, for any partition $\{I_1, I_2\}$ of the set $\{1, \dots, p\}$. This means that the set

$$SOL := \left\{ b \in \mathbb{R}^p : [(S_p)_{I_1} | (V_p)_{I_2}]^\top b = c^{I_1, I_2}, I_1 \cup I_2 = \{1, \dots, p\}, I_1 \cap I_2 = \emptyset \right\}$$

is finite, where $c_j^{I_1, I_2} := \sum_{i=1}^j (y_i - \lambda_i)$, for $j \in I_1$, and $c_j^{I_1, I_2} := 0$, for $j \in I_2$. Let b^* be any solution to (8.12). From Proposition 23 i), for all $j \in \{1, \dots, p\}$ we have $s_j^\top b^* = \sum_{i=1}^j (y_i - \lambda_i)$ or $v_j^\top b^* = 0$, which gives that $b^* \in SOL$. Since a feasible LP can have either one of infinitely many solutions, this immediately gives the claim. \blacksquare

Theorem 18 (Theorem 3.1 in [39]). *Consider an arbitrary linear program (LP) with objective $c^\top b$ and having an optimal solution (i.e., the optimal value is not $-\infty$) and let Q be a positive semidefinite matrix. Then there is a $\mu_0 > 0$ such that if $0 < \mu \leq \mu_0$, any solution to the perturbed problem with objective $c^\top b + \frac{1}{2}\mu(b - b_0)^\top Q(b - b_0)$ is a solution to LP. In particular, in the case where the LP solution is unique, the solution to the perturbed problem is unique and they coincide.*

Lemma 7. *Consider perturbed version of problem (8.12) with the same feasible set and objective function $f_\mu(b) := \lambda^\top b + \frac{1}{2}\mu\|b\|_2^2$ with $\mu > 0$. Let b^* be solution to perturbed problem (which is unique thanks to strong convexity). Then for any $y_1 \geq \dots \geq y_p \geq 0$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ (i.e. coefficients of λ does not have to be strictly decreasing and positive) it occurs $b_j^* = b_{j+1}^*$ or $\sum_{i=1}^j (y_i - \lambda_i) = \sum_{i=1}^j b_i^*$ for all $j \in \{1, \dots, p\}$, with the convention that $b_{p+1}^* := 0$ and $y_{p+1} := 0$.*

Proof. Take any feasible b and assume that for some $j \in \{1, \dots, p\}$ we have that $b_j > b_{j+1}$ and $\sum_{i=1}^j (y_i - \lambda_i) < \sum_{i=1}^j b_i$. Let b_ε be feasible vector constructed as in proof of Proposition 23 i). Then

$$\begin{aligned} f_\mu(b) - f_\mu(b_\varepsilon) &= \varepsilon(\lambda_j - \lambda_{j+1}) + \frac{1}{2}\mu(b_j^2 + b_{j+1}^2) - \frac{1}{2}\mu((b_j - \varepsilon)^2 + (b_{j+1} + \varepsilon)^2) = \\ &= \varepsilon(\lambda_j - \lambda_{j+1}) + \mu\varepsilon((b_j - b_{j+1}) - \varepsilon) > 0, \end{aligned}$$

for sufficiently small $\varepsilon > 0$. Hence, b can not be optimal. \blacksquare

Lemma 8. *Consider perturbed version of problem (8.12) as in Lemma 7, with objective function $f_\mu(b)$ and solution b^* . Moreover, assume that for some $j, l \in \{1, \dots, p\}$, $j < l$ we have $y_j - \lambda_j \leq y_{j+1} - \lambda_{j+1} \leq \dots \leq y_l - \lambda_l$. Let I_1 denote the set $\{j, \dots, l\}$ and $\overline{b_{I_1}}$ denote the arithmetic mean of subvector b_{I_1} , for any $b \in \mathbb{R}^p$. Then for any $y_1 \geq \dots \geq y_p \geq 0$ and $\lambda_1 \geq \dots \geq \lambda_p \geq 0$:*

- i) *solution is constant on the segment I_1 , i.e. $b_j^* = b_{j+1}^* = \dots = b_l^*$,*
- ii) *b^* is solution to perturbed problem with y and λ replaced respectively by \tilde{y} and $\tilde{\lambda}$,*

where

$$\tilde{\lambda}_i := \begin{cases} \overline{\lambda_{I_1}}, & i \in I_1 \\ \lambda_i, & \text{otherwise} \end{cases}, \quad \tilde{y}_i := \begin{cases} \overline{y_{I_1}}, & i \in I_1 \\ y_i, & \text{otherwise} \end{cases}. \quad (8.18)$$

Proof. To prove i), Suppose that $b_k^* > b_{k+1}^*$ for $k \in \{j, \dots, l-1\}$. Using the convention that $y_0 := b_0^* := \lambda_0 := 0$, from feasibility of b^* we have

$$\sum_{i=0}^{k-1} (y_i - \lambda_i) \leq \sum_{i=0}^{k-1} b_i^* \quad \text{and} \quad \sum_{i=0}^{k+1} (y_i - \lambda_i) \leq \sum_{i=0}^{k+1} b_i^*. \quad (8.19)$$

Adding both sides of these inequalities and dividing by 2 yields

$$\sum_{i=0}^{k-1} (y_i - \lambda_i) + \frac{(y_k - \lambda_k) + (y_{k+1} - \lambda_{k+1})}{2} \leq \sum_{i=0}^{k-1} b_i^* + \frac{b_k^* + b_{k+1}^*}{2} < \sum_{i=1}^k b_i^*. \quad (8.20)$$

Therefore

$$\sum_{i=1}^k (y_i - \lambda_i) \leq \sum_{i=0}^{k-1} (y_i - \lambda_i) + \frac{(y_k - \lambda_k) + (y_{k+1} - \lambda_{k+1})}{2} < \sum_{i=1}^k b_i^*, \quad (8.21)$$

which yields contradiction with Lemma 7.

To show that described modification of λ and y does not affect the solution, we will first show that feasible sets of both problems are identical. Let D and \tilde{D} denote the feasible sets for, respectively, initial parameters (y, λ) and $(\tilde{y}, \tilde{\lambda})$ given by (8.18). We start with proving that $\tilde{D} \subset D$. Let b be any vector from \tilde{D} . Since $\sum_{i=1}^k (y_i - \lambda_i) = \sum_{i=1}^k (\tilde{y}_i - \tilde{\lambda}_i)$, for $k < j$ and $k \geq l$, the task is reduces to showing that $\sum_{i=1}^k b_i \geq \sum_{i=1}^k (y_i - \lambda_i)$, for any $k \in \{j, \dots, l-1\}$. Since $\{y_i - \lambda_i\}_{i=j}^l$ increases, we simply have $\overline{y_{I_1}} - \overline{\lambda_{I_1}} \geq \frac{1}{k-j+1} \sum_{i=j}^k (y_i - \lambda_i)$. Using as before the convention $y_0 := \lambda_0 := 0$, we get

$$\sum_{i=1}^k b_i \geq \sum_{i=1}^k (\tilde{y}_i - \tilde{\lambda}_i) = \sum_{i=0}^{j-1} (y_i - \lambda_i) + (k-j+1) \cdot (\overline{y_{I_1}} - \overline{\lambda_{I_1}}) \geq \sum_{i=1}^k (y_i - \lambda_i). \quad (8.22)$$

Now, take any $b \in D$. After defining $b_0 := 0$ we have $(l-k) \cdot \sum_{i=0}^{j-1} b_i \geq (l-k) \cdot \sum_{i=0}^{j-1} (y_i - \lambda_i)$ and $(k-j+1) \cdot \sum_{i=0}^l b_i \geq (k-j+1) \cdot \sum_{i=0}^l (y_i - \lambda_i)$. Adding both sides of these inequalities and dividing by $(l-j+1)$ yields

$$\sum_{i=0}^{j-1} b_i + (k-j+1) \cdot \overline{b_{I_1}} \geq \sum_{i=0}^{j-1} (y_i - \lambda_i) + (k-j+1) \cdot (\overline{y_{I_1}} - \overline{\lambda_{I_1}}). \quad (8.23)$$

From the monotonicity of b , it occurs $\sum_{i=j}^k b_i \geq (k-j+1) \cdot \overline{b_{I_1}}$. Consequently,

$$\sum_{i=1}^k b_i \geq \sum_{i=0}^{j-1} b_i + (k-j+1) \cdot \overline{b_{I_1}} \geq \sum_{i=0}^{j-1} (y_i - \lambda_i) + (k-j+1) \cdot (\overline{y_{I_1}} - \overline{\lambda_{I_1}}) = \sum_{i=1}^k (\tilde{y}_i - \tilde{\lambda}_i) \quad (8.24)$$

and $D \subset \tilde{D}$ as a result.

Suppose now that b^* is solution for initial parameters (y, λ) , \tilde{b}^* is solution for $(\tilde{y}, \tilde{\lambda})$ and that

$$\frac{1}{2} \mu \|\tilde{b}^*\|_2^2 + \tilde{\lambda}^\top \tilde{b}^* < \frac{1}{2} \mu \|b^*\|_2^2 + \tilde{\lambda}^\top b^* \quad (8.25)$$

From i) we have $b_j^* = \dots = b_l^*$ and $\tilde{b}_j^* = \dots = \tilde{b}_l^*$, which yields $\tilde{\lambda}^\top \tilde{b}^* = \lambda^\top \tilde{b}^*$ and $\tilde{\lambda}^\top b^* = \lambda^\top b^*$. Therefore from (8.25) we have $f_\mu(\tilde{b}^*) < f_\mu(b^*)$, which contradicts the optimality of b^* . \blacksquare

proof of Theorem 17. Without loss of generality we can assume that we are starting with ordered and nonzero observations. Basing on Propositions 22 and 23, each solution to (ODS) is also a solution to (8.12). Since such solution is unique, this immediately gives the uniqueness of (ODS). Consider perturbed version of (8.12), with objective f_μ for sufficiently small μ , such as solutions to (8.12) and its perturbed version coincide (the existence of such μ is guaranteed by Theorem 18). Modifying y and λ as in the Procedure 7, after finite number of iterations we finish with converted y and λ such as

$$y_1 - \lambda_1 \geq \dots \geq y_p - \lambda_p. \quad (8.26)$$

From Lemma 8, we know that such modifications do not have an impact on the solution. Therefore, it is enough to show that, when assumption (8.26) is in use, the solution to SLOPE, $\beta^S = (y - \lambda)_+$, is also the unique solution, to (8.12).

With S_k and V_k defined in (8.17), the perturbed problem has following convex optimization form with affine inequality constraints

$$\begin{aligned} & \underset{b}{\text{minimize}} && \frac{1}{2}\mu\|b\|_2^2 + \lambda^\top b \\ & \text{s.t.} && \begin{cases} S_p^\top(y - \lambda - b) \leq 0, \\ -V_p^\top b \leq 0 \end{cases} \end{aligned} \quad (8.27)$$

If $y - \lambda \prec 0$, put $I_1 := \emptyset$, $I_2 := \{1, \dots, p\}$. Otherwise, let s be the maximal index such as $y_s - \lambda_s \geq 0$ and define $I_1 := \{1, \dots, s\}$, $I_2 := \{1, \dots, p\} \setminus I_1$. KKT conditions for (8.27) are given by

$$\mu b + \lambda = S_k \nu + V_k \tau, \quad (\text{Stationary})$$

$$\nu_i (S_p^\top(y - \lambda - b))_i = 0, \quad \tau_i (V_p^\top b)_i = 0, \quad \text{for each } i \in \{1, \dots, p\}, \quad (\text{Complementary slackness})$$

$$S_p^\top(y - \lambda - b) \leq 0, \quad -V_p^\top b \leq 0, \quad (\text{Primal feasibility})$$

$$\nu \succeq 0, \quad \tau \succeq 0. \quad (\text{Dual feasibility})$$

We will show that (b^*, ν^*, τ^*) satisfy KKT conditions, where $b^* = (y - \lambda)_+$ and ν^*, τ^* are given by

$$\nu_{I_1}^* := V_s^\top(\mu b^* + \lambda)_{I_1}, \quad \nu_{I_2}^* := 0, \quad \tau_{I_1}^* := 0, \quad \tau_{I_2}^* := S_{p-s}^\top(\mu b^* + \lambda)_{I_2}.$$

It is easy to see that b^* is primal feasible. Since coefficients of $\mu b^* + \lambda$ create nonnegative and nonincreasing sequence, we have $\nu^* \succeq 0$, $\tau^* \succeq 0$. Moreover, we easily get $(S_p)_{I_1}^\top(y - \lambda - b^*) = (S_s)^\top(y_{I_1} - \lambda_{I_1} - b_{I_1}^*) = 0$ and $(V_p)_{I_2}^\top b^* = (V_{p-s})^\top b_{I_2}^* = 0$, which shows that complementary slackness conditions are satisfied. Furthermore, we have

$$S\nu^* + V\tau^* = S_{I_1}\nu_{I_1}^* + V_{I_2}\tau_{I_2}^* = \begin{bmatrix} S_s & 0 \\ 0 & V_{p-s} \end{bmatrix} \begin{bmatrix} \nu_{I_1}^* \\ \tau_{I_2}^* \end{bmatrix} = \begin{bmatrix} (\mu b^* + \lambda)_{I_1} \\ (\mu b^* + \lambda)_{I_2} \end{bmatrix} = \mu b^* + \lambda, \quad (8.28)$$

which proves the stationary condition and finishes the proof. \blacksquare

We will now discuss the natural question about the generalization of Theorem 17 to an arbitrary norm g , i.e. whether the problems

$$\arg \min_b \left\{ \frac{1}{2}\|y - b\|_2^2 + g(b) \right\} \quad \text{and} \quad \begin{cases} \arg \min_b & g(b) \\ \text{s. t.} & g^D(y - b) \leq 1 \end{cases} \quad (8.29)$$

are equivalent for any norm g . We will show that, without additional assumptions concerning g , this is not true. As an counterexample consider $b \in \mathbb{R}^2$, $y := \begin{bmatrix} 0 \\ -2 \end{bmatrix}$ and $g(b) = \|Ab\|_1$, for $A := \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. Then, it holds

$$g^D(x) = \max_b \{x^\top b : \|Ab\|_1 \leq 1\} = \max_c \{x^\top A^{-1}c : \|c\|_1 \leq 1\} = \|(A^{-1})^\top x\|_1^D, \quad (8.30)$$

after substituting $c := Ab$. Hence, we get $g(b) = |b_1| + |b_1 + b_2|$ and $g^D(b) = \max\{|b_1 - b_2|, |b_2|\}$. We will now solve the left-hand side problem in (8.29), i.e. minimize $F(b) := \frac{1}{2}\|y - b\|_2^2 + |b_1| + |b_1 + b_2|$. Clearly, this problem has unique solution, b^* . Consider 4 sets, $S_1 := \{b : b_1 > 0, b_1 + b_2 > 0\}$, $S_2 := \{b : b_1 > 0, b_1 + b_2 < 0\}$, $S_3 := \{b : b_1 < 0, b_1 + b_2 > 0\}$, $S_4 := \{b : b_1 < 0, b_1 + b_2 < 0\}$. Function F is differentiable after restricting to S_i , for $i \in \{1, 2, 3, 4\}$, and it is easily to check that b^* does not belong to any of these sets. Therefore $b_1^* = 0$ or $b_1^* + b_2^* = 0$. After assuming the first case,

our problem reduces to minimize $\{\frac{1}{2}(b_2 + 2)^2 + |b_2|\}$, which simply gives $b^* = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$. The second case, in turn, yields problem minimize $\{\frac{1}{2}b_2^2 + \frac{1}{2}(b_2 + 2)^2 + |b_2|\}$ with solution $b^* = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$. Finally, comparing the values $F(0, -1) = \frac{3}{2}$ and $F(\frac{1}{2}, -\frac{1}{2}) = \frac{7}{4}$, we see that the former option constitutes the solution to starting problem.

Since $\|y - b^*\|_2^2 = 1$, the unconstrained problem in (8.29) is equivalent to minimizing function $g(b)$ over the set $\{b : \|y - b\|_2^2 = 1\}$. This problem was illustrated in Figure 11(a) and compare with the second problem from (8.29), Figure 11(b).

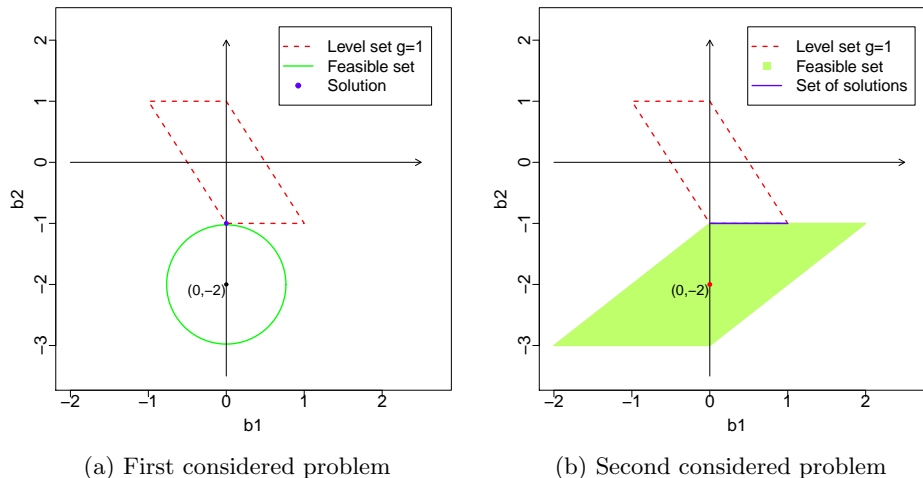


Figure 11: Comparison of solutions to both considered problems

Clearly, problems is (8.29) are not equivalent, since the right-hand problem has infinitely many solution, given by set $S^* = \{b \in \mathbb{R}^2, b_1 \in [0, 1], b_2 = -1\}$.

References

- [1] Tu Y.-K. and Kellett M. and Clerehugh V. and Gilthorpe M. S., “Problems of correlations between explanatory variables in multiple regression analyses in the dental literature,” *British Dental Journal*, vol. 199, pp. 457–461, 2005.
- [2] M. Bogdan, E. van den Berg, W. Su, and E. J. Candès, “Statistical Estimation and Testing via the Ordered ℓ_1 Norm,” *arXiv preprint at arXiv:1310.1969*, 2013.
- [3] H. Akaike, “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [4] G. Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
- [6] E. J. Candès and Y. Plan, “Near-ideal model selection by ℓ_1 minimization,” *Annals of Statistics*, vol. 37, no. 5A, pp. 2145–2178, 2009.

- [7] Hui Zou and Trevor Hastie, “Regularization and variable selection via the Elastic Net,” *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.
- [8] Hui Zou, “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [9] B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, pp. 407–499, 2004.
- [10] N. Meinshausen, “Relaxed lasso,” *Computational Statistics and Data Analysis*, pp. 374–393, 2007.
- [11] Emmanuel Candes and Terence Tao, “The Dantzig selector: Statistical estimation when p is much larger than n ,” *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.
- [12] 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*, vol. 57, no. 1, pp. 289–300, 1995.
- [13] R. Dutta, M. Bogdan, and J. K. Ghosh, “Model selection and multiple testing - a bayesian and empirical bayes overview and some new results,” *Journal of Indian Statistical Association*, pp. 105–142, 2012.
- [14] W. Su and E. Candès, “SLOPE is adaptive to unknown sparsity and asymptotically minimax,” *arXiv preprint arXiv:1503.08393*, 2015.
- [15] Bakin, S., “Adaptive regression and model selection in data mining problems,” 1999.
- [16] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society, Series B*, vol. 68, no. 1, pp. 49–67, 2006.
- [17] Julia E. Vogt and Volker Roth, “The Group-Lasso: $\ell_{1,\infty}$ Regularization versus $\ell_{1,2}$ Regularization,” *Proceedings of the 32nd DAGM conference on Pattern recognition*.
- [18] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group lasso,” *Journal of Computational and Graphical Statistics*, 2013.
- [19] N. Simon and R. Tibshirani, “Standardization and the group lasso penalty,” tech. rep., 2011.
- [20] J. Huang and T. Zhang, “The benefit of group sparsity,” tech. rep., 2009.
- [21] Karim Lounici and Massimiliano Pontil and Sara van de Geer, and Alexandre B. Tsybakov, “Oracle inequalities and optimal inference under group sparsity,” *The Annals of Statistics*, vol. 39, no. 4, pp. 2164–2204, 2011.
- [22] Lukas Meier and Sara van de Geer and Peter Bhlmann, “High-dimensional additive modeling,” *Annals of Statistics*, vol. 37, no. 6B, pp. 3779–3821, 2009.
- [23] Fengrong Wei and Jian Huang, “Consistent group selection in high-dimensional linear regression,” *Bernoulli*, vol. 16, no. 4, pp. 1369–1384, 2010.
- [24] P. Breheny and J. Huang, “Penalized methods for bi-level variable selection,” 2008.
- [25] Jian Huang and Patrick Breheny and Shuangge Ma, “A Selective Review of Group Selection in High-Dimensional Models,” *Statistical Science*, vol. 27, no. 4, pp. 481–499, 2012.
- [26] Seetharaman Indu, “Consistent bi-level variable selection via composite group bridge penalized regression,” tech. rep., 2013.

- [27] Jian Huang and Shuangge Ma, Huiliang Xie and Cun-Hui Zhang, “A Group Bridge Approach for Variable Selection,” tech. rep., 2007.
- [28] Boyd S. and Vandenberghe L., *Convex Optimization*. 2004.
- [29] Hardy, G.H. and Littlewood, J.E. and Pólya, G, *Inequalities*. 1952.
- [30] R. C. James, “Reflexivity and the supremum of linear functionals,” *Annals of Mathematics*, vol. 66, no. 1, pp. 159–169, 1957.
- [31] Nesterov Y., *Introductory Lectures on Convex Optimization. A Basic Course*. 1998.
- [32] Beck A. and Teboulle M., “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM Journal on Imaging Sciences*, vol. 1, pp. 183–202, 2009.
- [33] P. Tseng, “On accelerated proximal gradient methods for convex-concave optimization,” 2008.
- [34] A. Gossmann, S. Cao, and Y.-P. Wang, “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*, 2015.
- [35] R. J. Tibshirani, “The lasso problem and uniqueness,” 2012.
- [36] C. Sabatti, S. K. Service, A. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. J. Rvelin, N. B. Freimer, and L. Peltonen, “Genome-wide association analysis of metabolic traits in a birth cohort from a founder population,” *Nature Genetics*, vol. 41, no. 1, pp. 35–46, 2009.
- [37] J. S. C., “Hierarchical clustering schemes,” *Psychometrika*, vol. 32, pp. 241–254, 1967.
- [38] Langfelder, Peter and Zhang, Bin and Horvath, Steve, “Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R,” *Bioinformatics*, vol. 24, pp. 719–720, 2008.
- [39] S. Becker, E. J. Cands, and M. Grant, “Templates for convex cone problems with applications to sparse signal recovery,” 2010.

A gFDR control with gSLOPE under orthogonal case

A.1 The proof of Lemma 1

From Proposition 13 we know that $b_i^* > 0$ for $i \in \{1, \dots, r\}$. Let us define

$$\tilde{b}_i := \begin{cases} b_i^* - h, & i \in \{j, \dots, r\} \\ b_i^*, & \text{otherwise.} \end{cases},$$

where we restrict only to sufficiently small values of h , so as to the condition $\tilde{b}_i > 0$ is met for all i from $\{j, \dots, r\}$. For such h we have $b_{(r+1)}^* = \dots = b_{(p)}^* = \tilde{b}_{(r+1)} = \dots = \tilde{b}_{(p)} = 0$. Therefore there exists permutation $\pi : \{1, \dots, r\} \rightarrow \{1, \dots, r\}$ such as $\sum_{i=1}^r \lambda_i \tilde{b}_{(i)} = \sum_{i=1}^r \lambda_{\pi(i)} \tilde{b}_i$. For such permutation we have

$$\begin{aligned} J_\lambda(b^*) - J_\lambda(\tilde{b}) &= \sum_{i=1}^r \lambda_i b_{(i)}^* - \sum_{i=1}^r \lambda_i \tilde{b}_{(i)} = \sum_{i=1}^r \lambda_i b_{(i)}^* - \sum_{i=1}^r \lambda_{\pi(i)} \tilde{b}_i \geq \\ &\sum_{i=1}^r \lambda_{\pi(i)} b_i^* - \sum_{i=1}^r \lambda_{\pi(i)} \tilde{b}_i = h \sum_{i=j}^r \lambda_{\pi(i)} \geq h \sum_{i=j}^r \lambda_i, \end{aligned} \quad (\text{A.1})$$

where the first inequality follows from Theorem 2 and second is the consequence of monotonicity of $\{\lambda_i\}_{i=1}^p$. We also have

$$\begin{aligned} \|y - Db^*\|_2^2 - \|y - D\tilde{b}\|_2^2 &= \sum_{i=j}^r (y_i - d_i b_i^*)^2 - \sum_{i=j}^r (y_i - d_i b_i^* + d_i h)^2 = \\ &2h \sum_{i=j}^r (d_i^2 b_i^* - d_i y_i) - h^2 \sum_{i=j}^r d_i^2. \end{aligned} \quad (\text{A.2})$$

Optimality of b^* , (A.1) and (A.2) yield

$$0 \geq f(b^*) - f(\tilde{b}) \geq h \sum_{i=j}^r (d_i^2 b_i^* - d_i y_i + \lambda_i) - \frac{1}{2} h^2 \sum_{i=j}^r d_i^2, \quad (\text{A.3})$$

for each h from the interval $[0, \varepsilon]$, where $\varepsilon > 0$ is some (sufficiently small) value. This gives $\sum_{i=j}^r (d_i^2 b_i^* - d_i y_i + \lambda_i) \leq 0$ and consequently

$$\sum_{i=j}^r (d_i y_i - \lambda_i) \geq \sum_{i=j}^r d_i^2 b_i^* > 0. \quad (\text{A.4})$$

To prove claim (4.15), consider a new sequence defined as $\tilde{b}_i := \begin{cases} h, & i \in \{r+1, \dots, j\} \\ b_i^*, & \text{otherwise.} \end{cases}$. We will

restrict our attention only to $0 < h < \min\{b_i^* : i \leq r\}$, so as to $b_{(\cdot)}^*$ and $\tilde{b}_{(\cdot)}$ are given by applying the same permutation to b^* and \tilde{b} , respectively. Moreover, for each i from $\{r+1, \dots, j\}$ it holds $\tilde{b}_{(i)} = \tilde{b}_i = h$. From optimality of b^*

$$0 \geq f(b^*) - f(\tilde{b}) = \frac{1}{2} \sum_{i=r+1}^j (y_i^2 - (y_i - d_i h)^2) - \sum_{i=r+1}^j \lambda_i h = h \sum_{i=r+1}^j (d_i y_i - \lambda_i) - \frac{1}{2} h^2 \sum_{i=r+1}^j d_i^2,$$

for all considered h , which leads to (4.15).

A.2 The proof of Lemma 3

We have to show that solution \tilde{b}^* to problem

$$\underset{b}{\text{minimize}} \quad F(b) := \frac{1}{2} \sum_{i=1}^{p-1} (\tilde{y}_i - \tilde{d}_i b_i)^2 + \sum_{i=1}^{p-1} \tilde{\lambda}_i b_{(i)} \quad (\text{A.5})$$

has exactly $r - 1$ nonzero coefficients. From Proposition 11 we know that the change of signs of y_i 's does not affect the support, hence without loss of generality we can assume that $\tilde{y} \succeq 0$, and $\tilde{b}^* \succeq 0$ as a result (from Proposition 12). We will start with situation when $d_1 y_1 \geq \dots \geq d_p y_p$ and consequently $\tilde{d}_1 \tilde{y}_1 \geq \dots \geq \tilde{d}_{p-1} \tilde{y}_p$. If j is fixed index such as $d_j |y_j| > \lambda_r$ and $R(b^*) = r$, this gives

$$j \in \{1, \dots, r\}. \quad (\text{A.6})$$

To show that solution to (A.5) has at least $r - 1$ nonzero entries, for $r > 1$, suppose by contradiction that \tilde{b}^* has exactly $k - 1$ nonzero entries with $k < r$. Let us define $\hat{b} \in \mathbb{R}^{p-1}$ as

$$\hat{b}_i := \begin{cases} h, & i \in \{k, \dots, r-1\} \\ \tilde{b}_i^*, & \text{otherwise} \end{cases},$$

where $0 < h < \min\{\tilde{b}_1^*, \dots, \tilde{b}_{k-1}^*\}$. Then

$$F(\tilde{b}^*) - F(\hat{b}) = h \sum_{i=k}^{r-1} (\tilde{d}_i \tilde{y}_i - \tilde{\lambda}_i) - h^2 \sum_{i=k}^{r-1} \frac{1}{2} \tilde{d}_i^2. \quad (\text{A.7})$$

Now

$$\sum_{i=k}^{r-1} (\tilde{d}_i \tilde{y}_i - \tilde{\lambda}_i) = \sum_{i=k+1}^r (\tilde{d}_{i-1} \tilde{y}_{i-1} - \tilde{\lambda}_i) \geq \sum_{i=k+1}^r (d_i y_i - \lambda_i) > 0, \quad (\text{A.8})$$

where the first equality follows from $\tilde{\lambda}_i = \lambda_{i+1}$, the first inequality from $\tilde{d}_{i-1} \tilde{y}_{i-1} \geq d_i y_i$ and the second from Lemma 1. If h is small enough, we get $F(\hat{b}) < F(\tilde{b}^*)$ which leads to contradiction.

Suppose now by contradiction that \tilde{b}^* has k nonzero entries with $k \geq r$ and define

$$\hat{b}_i := \begin{cases} \tilde{b}_i^* - h, & i \in \{r, \dots, k\} \\ \tilde{b}_i^*, & \text{otherwise} \end{cases}.$$

Analogously to (A.1), we get $J_{\tilde{\lambda}}(\tilde{b}^*) - J_{\tilde{\lambda}}(\hat{b}) \geq h \sum_{i=r}^k \tilde{\lambda}_i$ and consequently

$$F(\tilde{b}^*) - F(\hat{b}) \geq h \left[\sum_{i=r}^k (\tilde{\lambda}_i - \tilde{d}_i \tilde{y}_i) + \sum_{i=r}^k \tilde{d}_i^2 \tilde{b}_i^* \right] - \frac{1}{2} h^2 \sum_{i=r}^k \tilde{d}_i^2. \quad (\text{A.9})$$

Now

$$\sum_{i=r}^k (\tilde{\lambda}_i - \tilde{d}_i \tilde{y}_i) = \sum_{i=r+1}^{k+1} (\lambda_i - d_i y_i) \geq 0, \quad (\text{A.10})$$

where the first equality follows from definition of $\tilde{\lambda}$ and (A.6), while the inequality follows from Lemma 1. If h is small enough, we get $F(\hat{b}) < F(\tilde{b}^*)$, which contradicts the optimality of \tilde{b}^* .

Consider now general situation, i.e. without assumption concerning the order of $D|y|$. Suppose that π , with corresponding matrix P_π , is permutation which orders $D|y|$. Define $y_\pi := P_\pi y$ and $D_\pi := P_\pi D P_\pi^\top$. Applying the procedure described in the statement of Lemma simultaneously to (y, D, λ) for j , and to (y_π, D_π, λ) for $\pi(j)$ we end with $(\tilde{y}, \tilde{D}, \tilde{\lambda}, \tilde{R}_1^j(\tilde{b}^*))$ and $(\tilde{y}_\pi, \tilde{D}_\pi, \tilde{\lambda}, \tilde{R}_2^{\pi(j)}(\tilde{b}_\pi^*))$. It is straightforward to see, that there exists permutation $\tilde{\pi} : \{1, \dots, p-1\} \rightarrow \{1, \dots, p-1\}$ such that $\tilde{y}_\pi = P_{\tilde{\pi}} \tilde{y}$ and $\tilde{D}_\pi = P_{\tilde{\pi}} \tilde{D} P_{\tilde{\pi}}^\top$. From Proposition 11 we have that $\tilde{b}_\pi^* = P_{\tilde{\pi}} \tilde{b}^*$ and $\tilde{R}_1^j(\tilde{b}^*) = \tilde{R}_2^{\pi(j)}(\tilde{b}_\pi^*)$. Moreover, from the first part of proof $\tilde{R}_2^{\pi(j)}(\tilde{b}_\pi^*) = r - 1$, which gives the claim.

B The expected values of random matrices

In this section of Appendix we will prove Lemmas 4 and 5.

B.1 The proof of Lemma 4

The first claim is obvious for $n = 1$ and we will assume that $n > 1$. First, we will list some basic properties of A_X . It could be easily noticed that: A_X is symmetric matrix, A_X is idempotent matrix (meaning that $A_X A_X = A_X$) and that $\text{trace}(A_X) = \text{trace}(X^T X (X^T X)^{-1}) = r$. We will now show that for each $i \in \{1, \dots, n\}$, $j \in \{1, \dots, n\}$ the support of a $A_X(i, j)$ distribution is bounded, which will give us the existence of the expected value. Indeed, from Theorem 11 we have

$$|(A_X)_{i,j}| \leq \|A\|_F = \sqrt{\text{trace}(A_X^T A_X)} = \sqrt{\text{trace}(A_X)} = \sqrt{r}. \quad (\text{B.1})$$

We will use notation $E_X := \mathbb{E}(A_X)$. Since entries of matrix X are randomized independently with the same distribution, E_X is invariant under permutation applied to rows, i.e. $E_X = E_{PX}$ for any permutation matrix P . This gives $E_X = P E_X P^T$, which means that applying the same permutation to rows and columns has no impact on expected value. We will show that

$$(E_X)_{i,j} = (E_X)_{1,n}, \text{ for } i < j. \quad (\text{B.2})$$

Consider first the case when $i = 1$ and $1 < j < n$. Denoting by $P_{j \leftrightarrow n}$ matrix corresponding to transposition which replaces elements j and n , we have $(E_X)_{1,j} = (P_{j \leftrightarrow n} E_X P_{j \leftrightarrow n}^T)_{1,j} = (E_X)_{1,n}$. When $j = n$ and $1 < i < n$, the same reasoning works with $P_{1 \leftrightarrow i}$. Suppose now, that $1 < i < n$ and $1 < j < n$. We get $(E_X)_{i,j} = (E_X)_{1,n}$ analogously by using arbitrary permutation matrix P which replaces element j with n and element i with 1 . Since A_X is symmetric, (B.2) is true also for $i > j$. On the other hand, for all $i, j \in \{1, \dots, n\}$, we have $(E_X)_{i,i} = (P_{j \leftrightarrow i} E_X P_{j \leftrightarrow i}^T)_{i,i} = (E_X)_{j,j}$. Consequently, all off-diagonal entries of E_X are equal to some t and all diagonal entries have the same value d . Since

$$nd = \text{trace}(E_X) = \sum_{i=1}^n \mathbb{E}(A_X(i, i)) = \mathbb{E}\left(\sum_{i=1}^n (A_X)_{i,i}\right) = r, \quad (\text{B.3})$$

we have $d = \frac{r}{n}$ and it remains to show that $t = 0$. Define $\Sigma := \begin{bmatrix} -1 & 0^T \\ 0 & \mathbf{I}_{n-1} \end{bmatrix}$. Then ΣX_S differs from X_S only by signs of the first row. Since entries of matrix X_S have zero-symmetric distribution, we have $E_X = E_{\Sigma X} = \Sigma E_X \Sigma$. Now

$$\begin{bmatrix} d & \mathbf{1}_{n-1}^T t \\ \mathbf{1}_{n-1} t & \ddots \end{bmatrix} = E_X = \Sigma E_X \Sigma = \begin{bmatrix} d & -\mathbf{1}_{n-1}^T t \\ -\mathbf{1}_{n-1} t & \ddots \end{bmatrix}, \quad (\text{B.4})$$

which implies $t = 0$ and proves the first lemma.

B.2 The proof of Lemma 5

Observe that $M_{X,\lambda}$ is symmetric, positive semi-definite matrix. In assumed situation we have $X^T X \sim \mathcal{W}_p(\frac{1}{n} \mathbf{I}_r, n)$, which together with Theorems 11 and 12 yields

$$\begin{aligned} \mathbb{E}|(M_{X,\lambda})_{i,j}| &\leq \mathbb{E}(\|M_{X,\lambda}\|_*) = \mathbb{E}(\text{trace}(M_{X,\lambda})) = \mathbb{E}(\text{trace}(H_{\lambda,\beta}^T B_X^T B_X H_{\lambda,\beta})) = \\ \mathbb{E}(H_{\lambda,\beta}^T (X^T X)^{-1} H_{\lambda,\beta}) &= \frac{n}{(n-r-1)} H_{\lambda,\beta}^T H_{\lambda,\beta} = \frac{n \|\lambda^S\|_2^2}{n-r-1}, \end{aligned} \quad (\text{B.5})$$

giving us the existence of $E_X := \mathbb{E}(M_{X,\lambda})$. Analogously to situation in Lemma 4, E_X is invariant under permutation or signs changes applied to rows of X , i.e. $E_X = E_{PX}$ for any permutation matrix P , and $E_X = E_{\Sigma X}$ for diagonal matrix Σ with entries on diagonal coming from set $\{-1, 1\}$. Since $E_{PX} = PE_XP^\top$ and $E_{\Sigma X} = \Sigma E_X \Sigma$, as before we have that E_X is diagonal matrix with all diagonal entries having the same value d . The value d could be easily found using (B.5) since we have

$$nd = \text{trace}(E_X) = \frac{n \|\lambda^S\|_2^2}{n - r - 1}. \quad (\text{B.6})$$

C Stopping criteria for numerical algorithm

Without loss of generality assume that $\sigma = 1$. We will start with optimization problem from (6.4), namely

$$\underset{\eta}{\text{minimize}} \quad f(\eta) = \frac{1}{2} \|y - XM\eta\|_2^2 + J_\lambda(\llbracket \eta \rrbracket_I), \quad (\text{C.1})$$

for $\llbracket \eta \rrbracket_I = (\|\eta_{I_1}\|_2, \dots, \|\eta_{I_m}\|_2)^\top$ and $M_{I_i, I_i} = \frac{1}{w_i} \mathbf{I}_{I_i}$, $i = 1, \dots, m$. This problem could be written in equivalent form

$$\begin{aligned} \underset{\eta, r, c}{\text{minimize}} \quad & \frac{1}{2} \|r\|_2^2 + c \\ \text{s.t.} \quad & \begin{cases} J_{\lambda, I}(\eta) - c \leq 0 \\ y - r - XM\eta = 0 \end{cases} \end{aligned} \quad (\text{C.2})$$

(notice that for (η^*, r^*, c^*) being solution, it must occur $c^* = J_{\lambda, I}(\eta^*)$). Since (C.2) is convex and (η_0, r_0, c_0) , for $\eta_0 = 0$, $r_0 = y$ and $c_0 = 1$, is strictly feasible, from Theorem 6 the strong duality holds. Lagrange dual function for this problem is given by

$$\begin{aligned} g(\mu, \nu) &= \inf_{\eta, r, c} \left\{ \frac{1}{2} \|r\|_2^2 + c + \mu^\top (y - r - XM\eta) + \nu (J_{\lambda, I}(\eta) - c) \right\} = \\ & \mu^\top y + \inf_r \left\{ \frac{1}{2} \|r\|_2^2 - \mu^\top r \right\} + \inf_c \{c - \nu c\} + \inf_\eta \{ -\mu^\top XM\eta + \nu J_{\lambda, I}(\eta) \}. \end{aligned} \quad (\text{C.3})$$

Now, since the minimum of $\frac{1}{2} \|r\|_2^2 - \mu^\top r$ is taken in $r = \mu$, we have

$$g(\mu, \nu) = \mu^\top y - \frac{1}{2} \|\mu\|_2^2 + \inf_c \{c - \nu c\} - J_{\nu\lambda, I}^*((XM)^\top \mu). \quad (\text{C.4})$$

From Proposition 17, the dual problem to (C.2) is therefore given by

$$\begin{aligned} \underset{\mu, \nu}{\text{maximize}} \quad & \mu^\top y - \frac{1}{2} \|\mu\|_2^2 \\ \text{s.t.} \quad & \begin{cases} \nu = 1 \\ \|(MX^\top \mu)_I\|_2 \in C_{\nu\lambda} \end{cases} \end{aligned} \quad (\text{C.5})$$

Let (η^*, r^*, c^*) be primal and (μ^*, ν^*) be dual solution to (C.2). Obviously, then we have $\nu^* = 1$, $\mu^* = r^* = y - XM\eta^*$ and $c^* = J_{\lambda, I}(\eta^*)$. Furthermore, from strong duality we have

$$\frac{1}{2} \|y - XM\eta^*\|_2^2 + J_{\lambda, I}(\eta^*) = (y - XM\eta^*)^\top y - \frac{1}{2} \|y - XM\eta^*\|_2^2, \quad (\text{C.6})$$

which gives $(XM\eta^*)^\top (y - XM\eta^*) = J_{\lambda, I}(\eta^*)$. Now, for current approximate $\eta^{[k]}$ of solution to (C.1), achieved after applying proximal gradient method, we define the current duality gap for k step as

$$\rho(\eta^{[k]}) = (XM\eta^{[k]})^\top (y - XM\eta^{[k]}) - J_{\lambda, I}(\eta^{[k]}) \quad (\text{C.7})$$

and we will determine the infeasibility of $\mu^{[k]} := y - XM\eta^{[k]}$ by using the measure

$$\text{infeas}(\mu^{[k]}) := \max \left\{ J_{\lambda, I}^D(MX^T \mu^{[k]}) - 1, 0 \right\} \quad (\text{C.8})$$

To define the stopping criteria we have applied the widely used procedure: treat $\rho(\eta^{[k]})$ as indicator telling how far $\eta^{[k]}$ is from true solution and terminate the algorithm when this difference and infeasibility measure are sufficiently small. Summarizing, we have derived algorithm according to scheme

Procedure 8 Numerical solver for gSLOPE

input: *infeas.tol: positive number determining the tolerance for infeasibility;*
dual.tol: positive number determining the tolerance for duality gap;
k := 0, $\eta^{[0]}$, $\mu^{[0]} := \mu(\eta^{[0]})$, $\text{infeas}^{[0]} := \text{infeas}(\mu^{[0]})$, $\rho^{[0]} := \rho(\eta^{[0]})$;

while ($\text{infeas}^{[k]} > \text{infeas.tol}$ or $\rho^{[k]} > \text{dual.tol}$) **do**

1. $k \leftarrow k + 1$;
2. get $\eta^{[k]}$ from Procedure 3;
3. $\mu^{[k]} := \mu(\eta^{[k]})$;
4. $\text{infeas}^{[k]} := \text{infeas}(\mu^{[k]})$, $\rho^{[k]} := \rho(\eta^{[k]})$;

end while
 $\beta^{\text{gS}} := M\eta^{[k]}$.
