SLOPE-Adaptive variable selection via convex optimization

Nafissa BENALI & Angel REYERO LOBO
Teacher: Guillermo DURAND

Institut Mathématique d'Orsay(IMO)
Paris-Saclay University

March 24, 2024

Abstract

To select the relevant variables in the sparse linear regression model, various approaches have been proposed in recent years. For instance, one of the most standard methods is the ℓ_1-norm penalization proposed by the Lasso. In this article, we present the *Sorted L-One Penalized Estimation* (SLOPE), introduced by Bogdan et al. (2015a), which aims to incorporate the adaptivity of the Benjamini-Hochberg (BH) procedure for variable selection in multiple testing into the ℓ_1 penalization framework. Specifically, it decouples it in the Lasso expression in the following sorted way

$$\min_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{[i]},$$

where $\lambda_1 \geq \ldots \geq \lambda_p$ and $|b|_{[1]} \geq \ldots \geq |b|_{[p]}$. Indeed, under the orthogonal design assumption, the False Discovery Rate (FDR) is provably controlled with a penalization sequence directly derived from the BH procedure. For a more general context, other penalization sequences are proposed. Certainly, the importance of this new convex optimization problem lies in its finite sample guarantees on the selected variables and its potential for generalization to other settings.

Keywords: Bonferroni correction, Benjamini-Hochberg procedure, False Discovery Rate, False Discovery Proportion, Sparse linear regression, Multiple testing, Lasso.

Notations. $\vee$ denotes the maximum between two numbers. Given a set $\mathcal{S}$, $|\mathcal{S}|$ is the cardinality of the set. $[\![a, b]\!]$ denotes the set of integers between a and b. Given a set $x_1, \ldots x_n \in \mathbb{R}$, we denote by $x_{(1)} \leq \ldots \leq x_{(n)}$ the set increasingly sorted and by $x_{[1]} \geq \ldots \geq x_{[n]}$ the set decreasingly sorted.

Contents

1 Introduction

This bibliographic work aims to summarize and extract the most important information presented in Bogdan et al. (2015a) and its online appendix Bogdan et al. (2015b).

1.1 Context

This project is undertaken as part of the evaluation for the subject *Guidelines in Machine Learning* in the second year of the Master's program in *Mathematics and Artificial Intelligence* at the Institut Mathématique d'Orsay (IMO) at Paris-Saclay University.

Additionally, we have created a corresponding Github repository containing the R code used to generate all the figures in this document. This code relies mainly on the R packages Tay et al. (2023) and Larsson et al. (2022). This repository aims to enhance the interpretability of the results.

1.2 Motivations

In the contemporary era, the expansion of data and the diverse methodologies employed for information extraction have presented statisticians new challenges distinct from traditional datasets. The objective now involves deriving insights from datasets where the number of features, denoted as p, may exceed that of samples n. A common reductive starting point for exploring relationships between covariates X and a target variable Y is the linear model, expressed as

$$Y = X\beta + Z,$$

where $Z \sim \mathcal{N}(0, \sigma^2 I_n)$ is normally distributed.

However, the primary aim of this work is not achieving optimal prediction accuracy but rather discerning the pertinent covariates for the target variable. Consider the following example: Suppose we have measured p genetic variants within a genomic region and seek to correlate these variants with a patient's cholesterol level. The study's purpose is not merely predicting cholesterol using genetic variants (directly measuring cholesterol would be more efficient) but rather offering insights for subsequent medical investigations. Then, our goal is to identify the relevant covariates. Discovering a relevant covariate could inform the development of cholesterol-related drugs or treatments. Otherwise, identifying irrelevant ones as relevant ones produces a wastage of resources.

In this sparse regression context, various methods grounded in convex minimization have emerged. One popular approach is the Lasso, which is an ℓ_1 convexification of model selection approach given by

$$\min_{b \in \mathbb{R}^p} \frac{1}{2} \|Y - Xb\|_2^2 + \lambda \|b\|_1. \tag{1}$$

As seen in Giraud (2021), under certain orthogonal assumptions on the design matrix, it yields satisfactory solutions. The problem with this kind of variable selection is that they do not provide finite sample guarantees on the chosen set. In fact, the penalisation parameter λ, usually selected by cross-validation, aims at maximising the accuracy of the prediction. However, as explained before, in our setting it is more important the validity of the selected variables than the prediction's accuracy.

To illustrate the concept of finite sample guarantees, let's consider an instance where a false positive occurs: when a variable is incorrectly identified as relevant. This leads to wasteful resource allocation. Consequently, it's crucial to minimize such false positives. A commonly used metric for assessing errors is the *Family Wise Error Rate* (FWER), which quantifies the probability of making at least one false positive.

The Bonferroni method, introduced by Bonferroni (1936), sets a threshold on p-values to control the FWER. Additionally, Holm (1979) proposed an adaptive iterative approach that refines the Bonferroni method, maintaining the same level of guarantee. However, these methods often are overly conservative, resulting in a lack of discoveries.

To address this issue, an alternative approach emerged: analyzing the proportion of false discoveries relative to the rejected set. Let $\mathcal{R}$ represent the rejected set, $\mathcal{H}_0$ denote the set of non-relevant indices, and $V(\mathcal{R}) = |\mathcal{R} \cap \mathcal{H}_0|$ indicate the cardinality of their intersection. Then, the False Discovery Proportion is defined as $V(\mathcal{R})/|\mathcal{R}| \vee 1$, with the False Discovery Rate(FDR) representing its expectation

$$\mathrm{FDR} \overset{\text{def}}{=} \mathbb{E}[\mathrm{FDP}] = \mathbb{E}\left[\frac{V(\mathcal{R})}{|\mathcal{R}| \vee 1}\right].$$

A popular approach to get this guarantee is to apply the step-up method proposed by Benjamini and Hochberg (1995). Given the p-values $p_1, \dots, p_p$ and a level $\alpha \in (0,1)$, we sort them $p_{(1)}, \dots, p_{(p)}$ in a increasing order with $p_{(0)} = 0$ by convention. Then, the recovered covariates will be the ones with p-value smaller than $\alpha \widehat{k}^{\mathrm{BH}}/p$ where

$$\widehat{k}^{\mathrm{BH}} = \max\left\{k \in [\![0, p]\!]; p_{(k)} \leq \frac{\alpha k}{p}\right\}. \tag{2}$$

We observe that a more stringent threshold is applied for smaller p-values, indicating stronger evidence to reject the null hypotheses. It is shown in Benjamini and Yekutieli (2001) that under some assumptions this method controls the FDR at a level α. These assumptions are generalized by a certain monotonicity concept: the weak Positive Regression Dependent on each from a Subset (wPRDS). In fact, all we need to know throughout this work is that for independent p-values and for positively correlated normal test statistics, this assumption is verified. For further information, we refer to Giraud (2021).

In the work conducted by Bogdan et al. (2015a), the authors attempt to adapt the previous methods to control the FDR in the sparse regression context. They decouple the ℓ_1 norm utilized in the Lasso (refer to Expression (1)) so that each covariate receives a distinct penalization based on its coefficient value. This approach follows the idea of BH of imposing a stricter penalization on hypotheses with stronger evidence for rejection. Additionally, penalizations in the orthogonal case are directly derived from the BH procedure, as it will be shown in Section 2.1.

In Section 2, we present the method valid under the orthogonal design matrix assumption. In Section 3, we provide further insights to generalize the penalization sequence. Finally, in Section 4, we discuss an example in which SLOPE demonstrates more desirable behavior compared to the Benjamini-Hochberg procedure, which can guide future research.

2 SLOPE

In this section we start by presenting the relationship between multiple testing and sparse regression in Section 2.1. In Section 2.2 we present the method, guarantees on the FDR control and an insight on practical implementation. Finally, in Section 2.3 we establish some common points and we compare it with other methods.

2.1 Intuitions with orthogonal design

In this section, we will simplify our scenario to develop an intuitive understanding of the thresholds we intend to employ. Specifically, we will examine the relationship between sorted penalization and multiple testing in a scenario where the design matrix X is assumed orthogonal and the noise is assumed to be i.i.d. with a known variance σ. In particular, we have

$$\widetilde{Y} \overset{\text{def}}{=} X^\top Y = X^\top X\beta + X^\top Z = \beta + X^\top Z \sim \mathcal{N}(\beta, \sigma^2 I_p).$$

Therefore, studying if the significancy of the β in the regression is similar to study a common multiple testing for the $\widetilde{Y}$. Indeed, for a coordinate $\widetilde{Y}_i \sim \mathcal{N}(\beta_i, \sigma^2)$, we can take the standard bilateral hypothesis testing which consist on rejecting the null hypothesis $\beta_i = 0$ if

$$\breve{p}_i = 2\left(1 - \Phi\left(\frac{|\widetilde{Y}_i|}{\sigma}\right)\right) < \alpha \text{ or equivalently } \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) < \frac{|\widetilde{Y}_i|}{\sigma},$$

where α is the confidence level and Φ (resp. Φ^{-1}) is the cumulative distribution function (resp. quantile function) of a standard Gaussian. However, this method does not work for the multiple testing setting. Therefore, as previously mentioned, it is possible to apply a Bonferroni correction (see Bonferroni (1936)) to control the FWER as follows:

$$\Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) < \frac{|\widetilde{Y}_i|}{\sigma}.$$

Moreover, we also note that under the orthogonal assumption, the solution of the Lasso minimization problem presented in Equation (1) is just a simple soft-thresholding where the estimate coefficient $\widehat{\beta}_j = 0$ if and only if $|\widetilde{Y}_j| > \lambda$. Hence, we can conceptualize the Lasso procedure with a penalization coefficient determined by the Bonferroni correction as $\lambda_{\text{Bonf}} = \sigma\Phi^{-1}(1 - \alpha/2p)$. This provides a control on the FWER.

Unfortunately, this approach is too conservative and lacks adaptability to the data. As illustrated in Figure 2, the penalization is so severe that it fails to make any significant discoveries. Additionally, we observe that this Lasso approach is not directly applicable to adaptive methods such as Holm (1979). One potential solution might involve iteratively reducing the threshold in a step-down procedure. However, this would necessitate multiple iterations of the Lasso procedure. While this should not pose a problem in sparse settings, even with this adaptivity, in cases of weak signals, it remains too restrictive as it fails to yield any discoveries with the initial threshold. Therefore, in the following, we recall the BH procedure presented in Equation (2):

$$\widehat{k}^{\text{BH}} = \max\left\{k \in [\![0, p]\!]; \breve{p}_{(k)} \leq \frac{\alpha k}{p}\right\} = \max\left\{k \in [\![0, p]\!]; \frac{|\widetilde{Y}_{[k]}|}{\sigma} \geq \Phi^{-1}\left(1 - \frac{k\alpha}{2p}\right)\right\}, \tag{3}$$

where the subscript (k) denotes the k-th smallest value of a sequence, while $[k]$ denotes the k-th greatest value. We observe that in this case the FDR is controlled at level α as the p-values are independent, then wPRDS (see Giraud (2021)). In fact, it is controlled at a level $\alpha p_0/p$ where $p_0 := |\mathcal{H}_0|$ is the number of null hypothesis.

2.2 Slope procedure

Building upon the intuitions gained from the previous section under the orthogonal design matrix, Bogdan et al. (2015a) propose a novel approach. They suggest a new optimization problem decoupling the ℓ_1-norm of the Lasso (1) into a sorted ℓ_1-norm that penalizes coefficients based on the BH procedure. Therefore, we present the method as follows

$$\min_{b\in\mathbb{R}^p} \frac{1}{2}\|Y - Xb\|_2^2 + \sum_{i=1}^{p} \lambda_i |b|_{[i]}, \tag{4}$$

where $\lambda_1 \geq \ldots \geq \lambda_p$ and $|b|_{[1]} \geq \ldots \geq |b|_{[p]}$. We first observe that the problem is convex. In fact, as demonstrated in Bogdan et al. (2015b), the penalization term is convex due to the use of an inequality (Hardy-Littlewood-Pólya), which establishes that it can be expressed as a maximum of convex functions, hence rendering it convex. Therefore, it is a tractable problem. Additionally, it is demonstrated in Bogdan et al. (2015b) that the penalization term constitutes a norm.

We note that while this method shares similarities with the BH procedure, they are not exactly the same, even in the orthogonal case, as we could find between the Lasso with the Bonferroni correction and the Bonferroni procedure. In fact, this method does not adhere to either a step-up or a step-down procedure with the BH thresholds. Abramovich and Benjamini (1995) proposed FDR-thresholding: a hard-thresholding method based on the BH procedure. However, we note that SLOPE has the advantage of being generalizable to contexts beyond the orthogonal context with another choice of λ (which will be presented in Section 3).

Guarantees of SLOPE: Although this method is not exactly equivalent to the BH method, it can be demonstrated that under certain assumptions, the FDR can be controlled.

Theorem 2.1 (SLOPE FDR control). *Under the linear assumption, orthogonal design matrix, noise $Z \sim \mathcal{N}(0, \sigma^2 I_n)$ with σ known, then the SLOPE procedure rejecting hypotheses for which $\widehat{\beta}_j \neq 0$ has an FDR controlled by*

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V(\mathcal{R})}{|\mathcal{R}| \vee 1}\right] \leq \alpha \frac{p_0}{p}.$$

The proof can be found on the online supplement of the article Bogdan et al. (2015b).

This theorem can be seen empirically in Figure 1 and in Figure 2. In fact, in both experiments, under the assumptions of the theorem, the FDR is controlled for different confidence levels α. Moreover, in Figure 2 we observe that the power, which is the proportion of true variables selected among the total number of true variables, is sufficiently large to ensure that we make discoveries.

Implementation: In practice, this optimization is tractable and there are multiple algorithms to approximate the solution of this problem. In Bogdan et al. (2015a), they proposed methods based on the proximity operator (prox). For instance, they proposed a general proximal gradient method based on approximating the sum of residuals by the second order Taylor expansion and iteratively applying a proximal mapping which can be computed efficiently. Moreover, similarly to the accelerations provided by the FISTA for the Lasso, they have proposed a more efficient version of the algorithm. Finally, they proposed another algorithm based on a reformulation of (4) into a quadratic program optimization problem which is then solved with a stack implementation.

2.3 Relation with other methods

Classical methods for variable selection are based on ℓ_0-norm penalizations. For instance, AIC, BIC or C_p de Mallow's. The problem with these methods is that the number of selected irrelevant variables tends to be large. In fact, they are usually based on developments that no longer hold for the high-dimensional setting, making the penalisation not strong enough. Other adaptive methods

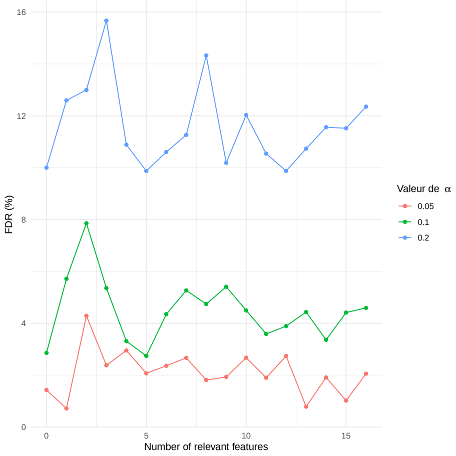

Figure 1: In this figure, we applied SLOPE with different confidence thresholds under an orthogonal design matrix. We observe that the FDR is controlled for all the levels α. The parameters used to create such graph are n = 800, p = 800, num_sims = 70, k_values = 0,...,16, var_error = 1, α = 0.05, 0.1, 0.2.

have been proposed based on this ℓ_0 penalizations (see Birgé and Massart (2001)). However, these approaches are not tractable.

A common tractable approach is to convexify the problem, leading to the Lasso. As shown in Equation (1), it depends on a penalization parameter λ. A standard method to choose it is based on cross-validation. However, as depicted in Figure 2, its main goal is to minimize the prediction error, but it is not able to control the FDR. Moreover, in the same figure, we observe that choosing λ to control the FWER following the Bonferroni correction leads to zero discoveries

Numerous variants of the Lasso have been proposed. For example, the Adaptive-Lasso, introduced by Zou (2006), adjusts the penalty for each component individually. Specifically, it employs a weighted ℓ_1-norm, where the weights are inversely proportional to the magnitudes of the initial coefficient estimates. Consequently, coefficients with smaller initial estimates receive higher penalties. This approach contrasts with SLOPE, where larger coefficients are penalized more heavily. What we can see in Figure 2 is that there's no guarantee on the selected variables, which is why the False Discovery Rate (FDR) of the adaptive lasso method is poor. However, the power is good since there's no variable selection. Mean Squared Error (MSE) is also good because it's data-driven to achieve proper prediction.

Correction: The introduced penalization imposes a bias on the Lasso estimate. One approach to mitigate this bias is the LSLasso (Least Squares Lasso), as outlined in Lederer (2013). The LSLasso operates through a two-step algorithm: initially, a Lasso regression is executed to select relevant coefficients, followed by a Least Squares regression performed only on the identified support set. A comparable strategy is suggested for debiasing the SLOPE solution.

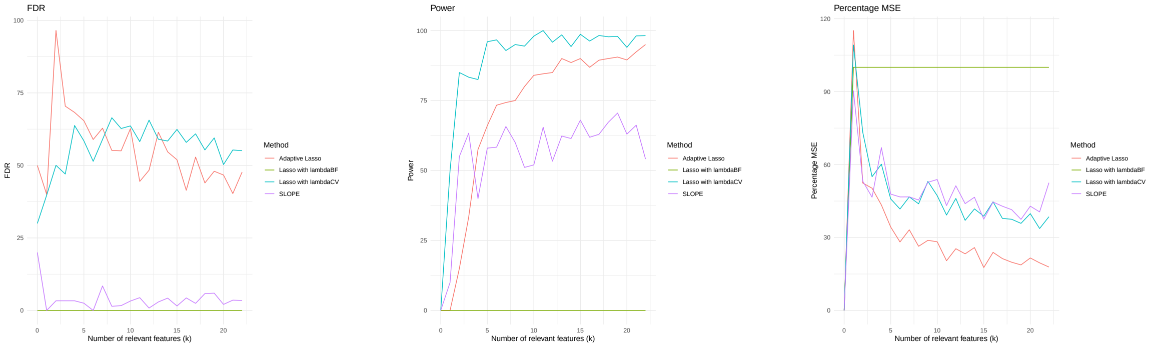

Figure 2: We compare the SLOPE procedure with different Lasso variants (with penalization term chosen by Bonferroni correction or by CV, and Adaptive-Lasso) under an orthogonal design matrix. On the left, we observe that SLOPE controls the FDR, while Lasso-λ_{CV} explodes. Lasso-λ_{Bonf} controls the FDR because it does not make any discoveries, as seen in the center. SLOPE achieves good power even while controlling the FDR, while Lasso-λ_{CV} achieves better power by accepting too many parameters. On the right, as expected, we observe the good performance of Lasso-λ_{CV} for prediction, while SLOPE does not achieve this performance because it does not select all the relevant variables. The parameters used to create such graphs are n = 500, p = 500, num_sims = 10, k_values = 0,...,30, var_error = 1, α = 0.1.

3 Towards a general SLOPE

All the previous discussion treated orthogonal design matrices. To make the algorithm more general, Bogdan et al. (2015a) proposed certain insights to modify the penalization sequence, considering the variance introduced by non-orthogonal design matrices. Hence, in this section, we present mathematical developments that are far from constituting rigorous proofs validating the selection of this new parameter λ for the minimization problem (4), but that proves satisfying empirical results.

As usual, we start with the Lasso to gain an insight into the effect of shrinkage due to penalization. We assume unit norm on the columns of X and that the noise z is distributed as a standard Gaussian. Then, the solution of the Lasso satisfies

$$\widehat{\beta} = \eta_\lambda(\widehat{\beta} - X^\top(X\widehat{\beta} - Y)) = \eta_\lambda(\widehat{\beta} - X^\top X(\widehat{\beta} - \beta) + X^\top z),$$

with η_λ the soft-thresholding operator $\eta_\lambda(t) = \text{sign}(t)(|t| - \lambda)_+$. Defining $v_i := \langle X_i, \sum_{i \neq j} X_j(\beta_j - \widehat{\beta}_j)\rangle$ we can rewrite the above condition as $\widehat{\beta}_i = \eta_\lambda(\beta_i + X_i^\top z + v_i)$. In the orthogonal case, it would be equal to zero so that $\widehat{\beta}_i = \eta_\lambda(\beta_i + X_i^\top z)$ with $X_i^\top z \sim \mathcal{N}(0,1)$ conditionally to X. Then, in this case it would be enough to apply a Bonferroni correction with a λ such that $\mathbb{P}(\max_i |X_i^\top z| > \lambda) \leq \alpha$. When this is not the case, the v_i does not vanish and in fact it increases with the estimation error of the β.

The aim of the method is to iteratively consider the shrinkage induced by the already selected coefficients with their respective penalizations. To achieve this, Bogdan et al. (2015a) suggested multiplying the standard λ_{BH} by a variance corrector based on the prior penalizations. We make the assumption that the selected coefficients for the estimate are exactly those from the support of β, defined by $\mathcal{S}$. Therefore, we would like an estimate of $X_i^\top X_{\mathcal{S}}(\beta_{\mathcal{S}} - \widehat{\beta}_{\mathcal{S}})$. First we note that under the assumption of having all the $\beta \geq 0$, the KKT conditions for the Lasso gives $X_{\mathcal{S}}^\top(Y - X\widehat{\beta}_{\mathcal{S}}) = \lambda \mathbf{1}_{\mathcal{S}}$, or equivalently, $\widehat{\beta}_{\mathcal{S}} = (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}(X_{\mathcal{S}}^\top Y - \lambda \mathbf{1}_{\mathcal{S}})$. Similarly for the SLOPE, we get $\widehat{\beta}_{\mathcal{S}} = (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}(X_{\mathcal{S}}^\top Y - \lambda_{\mathcal{S}})$ where the penalization term depends on the coordinate. We observe that the first term is the unbiased classical Least Square estimate, so that conditionally on the design matrix we obtain $\mathbb{E}[\beta_{\mathcal{S}} - \widehat{\beta}_{\mathcal{S}}] \approx (X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\lambda_{\mathcal{S}}$, i.e. $\mathbb{E}[X_i^\top X_{\mathcal{S}}(\beta_{\mathcal{S}} - \widehat{\beta}_{\mathcal{S}})] \approx X_i^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\lambda_{\mathcal{S}}$ for $i \notin \mathcal{S}$. Finally we make the assumption that the design matrix is Gaussian $\mathcal{N}(0, 1/n)$ and we obtain that $\mathbb{E}[(X_i^\top X_{\mathcal{S}}(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}\lambda_{\mathcal{S}})^2] = \lambda_{\mathcal{S}}^\top \mathbb{E}[(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1}]\lambda_{\mathcal{S}}/n = w(|\mathcal{S}|)\|\lambda_{\mathcal{S}}\|_2^2$, where $w(k) = 1/(n-k-1)$. This latter result is based on the expectation of an inverse Wishart. When the Gaussian assumption is not made, Monte Carlo can be used to estimate the correction.

Given these insights, the corrected penalization sequence λ_G (denoted as G due to the Gaussian assumption) is described as having the same initial penalization $\lambda_G(1) = \lambda_{\mathrm{BH}}(1)$. For the second step, we incorporate the correction term by $\lambda_G(2) = \lambda_{\mathrm{BH}}(2)\sqrt{1 + w(1)\lambda_G(1)^2}$. Similarly, we proceed by updating $\lambda_G(i) = \lambda_{\mathrm{BH}}(i)\sqrt{1 + w(i-1)\sum_{j<i}\lambda_G(j)^2}$. It is important to note that this sequence is not necessarily decreasing. However, to maintain the convexity of the penalization, it must be decreasing. Therefore, if there is an index where the penalization is greater than the previous one, we set the subsequent values in the sequence to be equal to the previous value.

We emphasize that this heuristic method is far from formal. However, it often yields good empirical results. Therefore, future work could focus on establishing some properties of this sequence.

4 Is SLOPE better than BH procedure?

In this section, we provide an intuitive explanation using an example given in Bogdan et al. (2015a) of a positively correlated normal test where the SLOPE procedure exhibits more desirable properties compared to the BH procedure. In this experience we suppose that $p = 1000$ experiments have been done in 5 laboratories where each observation is modeled as

$$y_{i,j} = \mu_i + \tau_j + z_{i,j} \text{ for } 1 \leq i \leq 1000, 1 \leq j \leq 5,$$

where the laboratory effects $\tau_j \sim \mathcal{N}(0, \sigma_\tau^2)$ are i.i.d. and independent from the errors $z_{i,j} \sim \mathcal{N}(0, \sigma_z^2)$ which are also i.i.d. The first approach presented is a multiple testing procedure. Indeed, taking the mean over the five laboratories, we have

$$\overline{y}_i = \mu_i + \overline{\tau} + \overline{z}_i \text{ for } 1 \leq i \leq 1000,$$

so that $\overline{y} \sim \mathcal{N}(\mu, \Sigma)$ with $\sigma := \Sigma_{i,i} = (\sigma_\tau^2 + \sigma_z^2)/5$ and $\rho := \Sigma_{i,j} = \sigma_\tau^2/5$. We observe that all the entries of the matrix Σ are positive. Consequently, we are in the scenario of positively correlated normal test statistics, where standard bilateral tests satisfy the wPRDS property (as discussed in Chapter 10 of Giraud (2021)), ensuring that the BH procedure controls the FDR. Then, we can apply it by first ordering $|\overline{y}|_{[1]} \geq \ldots \geq |\overline{y}|_{[p]}$ and then apply a Step-Up procedure with $\sigma\Phi^{-1}(1 - i\alpha/2p)$ for the i-th biggest value.

Another possible approach simply consists of whitening the noise to transform the problem into a regression one, i.e.

$$\widetilde{y} = \Sigma^{-1/2}\overline{y} = \Sigma^{-1/2}\mu + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, I_p).$$

Then, by considering $\Sigma^{-1/2}$ as the design matrix, it is possible to apply SLOPE to recover the non-null coefficients of μ, which is the objective of the study. First note that $\Sigma^{-1/2}$ is not an orthogonal design matrix therefore Theorem 2.1 no longer applies. However, it can be proven to be diagonally dominant (for example, for $\sigma_\tau^2 = \sigma_z^2 = 2.5$ we have $\Sigma_{i,i}^{-1/2} = 1.4218$ and $\Sigma_{i,j}^{-1/2} = -0.0014$) so approximately orthogonal design.

In a more general context, the error variance is unknown. It is for this reason that Bogdan et al. (2015a) have proposed some estimates based on classical unweighted means method. For the ease of message that we want to transfer, we will continue with the known variance assumption. In fact, we retake the experiences from Bogdan et al. (2015a) with $\sigma_\tau^2 = \sigma_z^2 = 2.5$ and the nonzero set to $\sqrt{2\text{log}p}/c$ with c the Euclidean norm of each of the columns of $\Sigma^{-1/2}$.

As seen in Figure 3, the FDR is controlled for both methods. However, as seen in Figure 4, the FDP of the BH procedure is either concentrated at zero or distributed uniformly between $[0, 1]$. This behavior is undesirable because it usually either makes no discoveries, or when it does, there are no guarantees on them. Indeed, we observe that as the FDR is a sort of mean of the previous, the non-discovery simulations compensate the others and then control the FDR. In contrast, SLOPE ensures guarantees whenever a variable is selected. This was also observed in the good power observed in Figure 2.

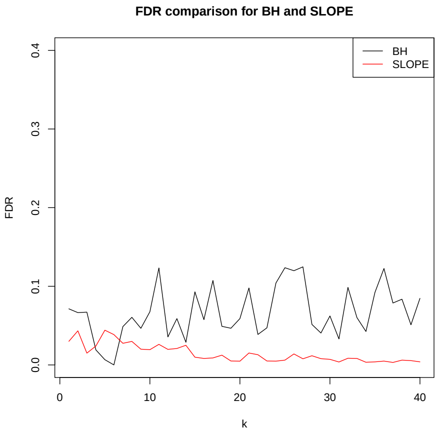

Figure 3: We observe that FDR is controlled for both methods. The parameters used to create such graph are I = 1000, J = 5, $\alpha = 0.1$, $\sigma_\tau^2 = 2.5$, $\sigma_z^2 = 2.5$, $\sigma = \frac{\sigma_\tau^2 + \sigma_z^2}{J}$, k_values = 1,...,40.

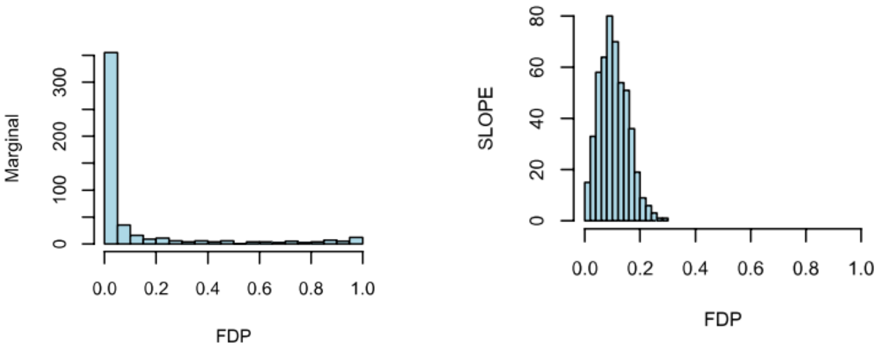

Figure 4: (Figure from Bogdan et al. (2015a)) Histograms of 500 simulations with $k = 50$ and $q = 0.1$. On the left, we observe that the FDPs of the BH-procedure are either concentrated around zero or uniformly distributed across the entire interval. On the right, the FDPs of SLOPE are more evenly distributed, providing greater assurance regarding the rejections.

5 Conclusion & Perspectives

Bogdan et al. (2015a) proposed a tractable convex optimization problem for variable selection, which is a kind of mixture of Lasso and the BH procedure. The FDR is provably controlled under the orthogonal design matrix. One of the advantages of this method compared to the literature is its ability to be used in a more general context. Moreover, they have even proposed a general sequence for penalization for more general sets.

Another significant advantage of this method is that, as seen in Figure 4, there may be some control over the FDP instead of only over the FDR. This is a particularly powerful result as it can provide better guarantees on the discoveries made by the method. Therefore, future work may focus on establishing assumptions to further guarantee this result.

Another aspect to consider is formalizing the reasoning presented in Section 3 to develop a more robust general sequence with theoretical guarantees.

Finally, we would like to conclude by highlighting that all the experiments except the last one were conducted using different data from the article. Specifically, the numerical experiments described in the article were conducted using correlated design matrices distributed by $\mathcal{N}(0, 1/nI_p)$. Instead of using the BH penalization sequence, they utilized the generalized one. We though that this was *tricky* because they employed it to empirically demonstrate the FDR control of the method when they were not usen the provably method. Therefore, we opted to use the BH penalization sequence under conditions of orthogonality of the design.

Moreover, as noted in Section 3, the sequence was specifically constructed under the assumption of a $\mathcal{N}(0, 1/nI_p)$ design matrix, which aligns with the data used in their experiments and the implementation in their R package. Thus, one may question whether the method is as general as they claim.

References

Abramovich, F. and Benjamini, Y. (1995). *Thresholding of Wavelet Coefficients as Multiple Hypotheses Testing Procedure*, pages 5–14. Springer New York, New York, NY.

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

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268.

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

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015b). Supplement to "SLOPE—Adaptive variable selection via convex optimization". `https://doi.org/10.1214/15-AOAS842SUPP`. Online Appendix containing proofs of some technical results discussed in the text.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. 8:3–62.

Giraud, C. (2021). *Introduction to high-dimensional statistics.* CRC Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Larsson, J., Wallin, J., Bogdan, M., van den Berg, E., Sabatti, C., Candes, E., Patterson, E., Su, W., Kała, J., Grzesiak, K., and Burdukiewicz, M. (2022). *SLOPE: Sorted L1 Penalized Estimation.* R package version 0.5.0.

Lederer, J. (2013). Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.