Stepdown SLOPE for Controlled Feature Selection

Jingxuan Liang[1], Xuelin Zhang [2], Hong Chen[1, 4, 6,*], Weifu Li[1, 5, 6], Xin Tang[3]
[1]College of Science, Huazhong Agricultural University, Wuhan 430070, China
[2] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[3]Ping An Property & Casualty Insurance Company, Shenzhen, China
[4]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[5]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070, China
[6]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan 430070, China
chenh@mail.hzau.edu.cn

Abstract

Sorted L-One Penalized Estimation (SLOPE) has shown the nice theoretical property as well as empirical behavior recently on the false discovery rate (FDR) control of high-dimensional feature selection by adaptively imposing the non-increasing sequence of tuning parameters on the sorted ℓ_1 penalties. This paper goes beyond the previous concern limited to the FDR control by considering the stepdown-based SLOPE to control the probability of k or more false rejections (k-FWER) and the false discovery proportion (FDP). Two new SLOPEs, called k-SLOPE and F-SLOPE, are proposed to realize k-FWER and FDP control respectively, where the stepdown procedure is injected into the SLOPE scheme. For the proposed stepdown SLOPEs, we establish their theoretical guarantees on controlling k-FWER and FDP under the orthogonal design setting, and also provide an intuitive guideline for the choice of regularization parameter sequence in much general setting. Empirical evaluations on simulated data validate the effectiveness of our approaches on controlled feature selection and support our theoretical findings.

Introduction

Feature selection aims to find the informative features from high-dimensional empirical observations, which is one of key research fields of machine learning. Typical feature selection methods include sparse linear models (e.g., Lasso (Tibshirani 1996)), sparse additive models (e.g., SpAM (Ravikumar et al. 2009), GroupSAM (Chen et al. 2017), SpMAM (Chen et al. 2021)), tree-based models (e.g., random forest (Breiman 2001)), and sparse neural networks (e.g., LassoNet (Lemhadri, Ruan, and Tibshirani 2021)).

Following this line, the controlled feature selection further addresses the selection quality with low false discovery rate (FDR) guarantee, which has attracted the increasing attention recently due to its wide applications, e.g., in bioinformatics and biomedical (Aggarwal and Yadav 2016; Yu, Kaufmann, and Lederer 2021). There are mainly three branches of learning systems for controlled feature selection: the multiple hypothesis test (Benjamini and Hochberg 1995; Ferreira and Zwinderman 2006; Lehmann and Romano 2005; Romano and Shaikh 2006), the knockoffs filter (Barber and Candès 2015; Candès et al. 2018; Barber, Candès, and Samworth 2020; Romano, Sesia, and Candès 2020), and the Sorted L-One Penalized Estimation (SLOPE) (Bogdan et al. 2015; Su and Candès 2016; Brzyski et al. 2019). As a classic strategy for feature selection, the Benjamini and Hochberg (BH) procedure is formulated by jointly considering p-values of multiple hypothesis testing. Despite this procedure enjoys nice theoretical properties on the FDR control, it may face the computation challenge for nonlinear and complex regression estimation (Javanmard and Javadi 2019). As a novel feature filter scheme, the knockoffs inference has solid theoretical foundations and shows the competitive performance in real-word applications (Barber and Candès 2015; Barber, Candès, and Samworth 2020; Zhao et al. 2022; Yu, Kaufmann, and Lederer 2021). Particularly, an error-based knockoffs inference framework is formulated in (Zhao et al. 2022) to further realize the controlled feature selection from the perspectives of the probability of k or more false rejections (k-FWER) and the false discovery proportion (FDP). Different from screening out the active feature with the help of knockoff features, SLOPE focuses on the regularization design for sparse feature selection, which adaptively imposes a non-increasing sequence of tuning parameters on the sorted ℓ_1 penalties (Bogdan et al. 2015; Brzyski et al. 2019; Jiang et al. 2022).

Although rapid progresses on its optimization algorithm (Bogdan et al. 2015; Brzyski et al. 2019) and theoretical properties (Su and Candès 2016), all the existing works of SLOPE are limited to the FDR control only. Naturally, it is important to explore new SLOPE for controlled feature selection under other statistical criterion, e.g., k-FWER and FDP.

To fill this gap, we propose new SLOPE approaches, called k-SLOPE and F-SLOPE, to realize feature selection with the k-FWER and FDP control respectively. Different from the previous method relying on BH procedure, the proposed SLOPEs depend on the stepdown procedure (Lehmann and Romano 2005), which enjoy much feasibility and adpativity (Bogdan et al. 2015; Su and Candès 2016). The main contributions of this paper are summarized as below:

- *New SLOPEs for the kFWER and the FDP control.* We integrate the SLOPE (Bogdan et al. 2015) and the step-

[*] Corresponding author.

down procedure (Lehmann and Romano 2005) into a coherent way for the k-FWER and FDP control and formulate the respective convex optimization problem. Similarly with the flexible knockoffs inference in (Zhao et al. 2022), our approaches also can avoid the complex p-value calculation and can be implemented feasibility.

- *Theoretical guarantees and empirical effectiveness*. Under the orthogonal design setting, the k-FWER and FDP can be provably controlled at a prespecified level for the proposed k-SLOPE and F-SLOPE, respectively. In addition, we provide an intuitive theoretical analysis for the choice of the regularizing sequence in general setting. Simulated experiments validate the effectiveness of our SLOPEs on the k-FWER and FDP control, and verify our theoretical findings.

Related Work

To better highlight the novelty of the proposed method, we review the related SLOPE methods as well as the relationship among FDR, k-FWER and FDP.

SLOPE Methods. SLOPE (Bogdan et al. 2015) can be considered as a natural extension of Lasso (Tibshirani 1996), where the regression coefficients are penalized according to their rank. One notable choice of the regularization sequence $\{\lambda_i\}$ is given by the BH (Benjamini and Hochberg 1995) critical values $\lambda_{\mathrm{BH}}(i) = \Phi^{-1}(1 - \frac{iq}{2m})$, where $q \in (0,1)$ is the desired FDR level, m is the characteristic number and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. The main motivation behind SLOPE is to provide finite sample guarantees on regression estimation and FDR control, where FDR is defined as the expected proportion of irrelevant regressors among all selected predictors. When X is an orthogonal matrix, SLOPE with λ_{BH} controls FDR at the desired level in theory. Besides, a remarkable feature is that SLOPE does not require any knowledge of the degree of sparsity, yet automatically yields optimal total squared errors over a wide range of ℓ_0-sparsity classes.

To improve computing efficiency, a sparse semismooth Newton-based augmented Lagrangian technique was proposed to solve the more general SLOPE model (Luo et al. 2019). A heuristic screening rule for SLOPE based on the strong rule for the lasso was first presented in order to improve the numerical procedures efficiency of SLOPE, especially in the setting of estimating a complete regularization path (Larsson, Bogdan, and Wallin 2020). And Larsson et al. (2022) also proposed a new fast algorithm to solve the SLOPE optimization problem, which combined proximal gradient descent and proximal coordinate descent steps. Besides the above works on algorithm optimization, there are extensive studies on SLOPE with properties (Su and Candès 2016; Bellec, Lecué, and Tsybakov 2018; Kos and Bogdan 2020), model improvements (Brzyski et al. 2019; Lee, Sobczyk, and Bogdan 2019; Riccobello et al. 2022; Jiang et al. 2022) and applications (Brzyski et al. 2017; Kremer et al. 2020). As we know, there is no any touch to address the SLOPE-based feature selection with k-FWER or FDP control guarantees.

Statistical Metrics: FDR, k-FWER and FDP. Benjamini and Hochberg (1995) formulated the BH procedure to the control the expectations of FDP, called FDR control. Then, Lehmann and Romano (2005) proposed both the single step procedure and the stepdown procedure in order to ensure the k-FWER control. Lehmann and Romano (2005) also considered the FDP control and provided two stepdown procedures for controlling the FDP under mild conditions with the p-values dependence structure or no any dependence supposition. With the help of stepdown procedures (Lehmann and Romano 2005), there are studies on feature selection with the k-FWER control (Romano and Shaikh 2006; Romano and Wolf 2007; Alemán et al. 2017; Zhao et al. 2022) and the FDP control (Romano and Shaikh 2006; Romano and Wolf 2007; Fan and Lv 2010; Delattre and Roquain 2015; Zhao et al. 2022). However, most of these procedures may depend on the p-values to assess the importance of each feature or the assumption of structures. Moreover, the traditional calculation of p-value relies on the large-sample asymptotic theory usually, which may no longer be true in the setting of high-dimensional finite samples (Candès et al. 2018; Fan, Demirkaya, and Lv 2019).

It is necessary to explain the relationship between FDR, FDP and k-FWER. Given $\gamma, \alpha \in (0,1)$, the FDP control means the $\mathrm{Prob}(\mathrm{FDP} > \gamma)$ at the level α. Recall that the FDP concerns

$$\mathrm{Prob}\{\mathrm{FDP} > \gamma\} < \alpha, \tag{1}$$

and FDR is the expectation of FDP, i.e., $\mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$. It is easy to verify that

$$\mathrm{FDR} \leq \gamma \mathrm{Prob}\{\mathrm{FDP} \leq \gamma\} + \mathrm{Prob}\{\mathrm{FDP} > \gamma\},$$

and then

$$\frac{\mathrm{FDR} - \gamma}{1 - \gamma} \leq \mathrm{Prob}\{\mathrm{FDP} > \gamma\} \leq \frac{\mathrm{FDR}}{\gamma},$$

where the last inequality follows from the Markov's inequality. Clearly, if a method controls FDR at level q, then it also controls $\mathrm{FDP} \leq q/\gamma$. Conversely, if the FDP is controlled, i.e. $\mathrm{Prob}(\mathrm{FDP} > \gamma) < \alpha$, and then the FDR is bounded by $(1-\gamma)\alpha + \gamma$. Therefore, a procedure with the FDP control often can control the FDR (Van der Laan, Dudoit, and Pollard 2004). Furthermore, Farcomeni (2008) pointed out that, compared with the FDR control, the k-FWER control is more desirable when powerful selection results can be made.

Preliminaries

This section recalls some necessary backgrounds involved in this paper, e.g., SLOPE (Bogdan et al. 2015) and the stepdown procedure (Lehmann and Romano 2005).

Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^m$ and $\mathcal{Y} \subset \mathbb{R}$ be the compact input space and corresponding output space, respectively. Consider samples $\{(x_i, y_i)\}_{i=1}^n$ independently drawn from an unknown distribution on $\mathcal{X} \times \mathcal{Y}$. Denote

$$X := (X_1, X_2, \cdots, X_n)^T \subset \mathbb{R}^{n \times m}$$

and

$$y = (y_1, y_2, \cdots, y_n)^T \in \mathbb{R}^n.$$

The module length of each column vector of X is equal to 1. The output vector y is generated by the following multiple linear regression model:

$$y = X\beta + \epsilon, \tag{2}$$

where $\beta \in \mathbb{R}^m$ represents the coefficient vector and $\epsilon \sim N(0, \sigma^2 I_n)$. In sparse high-dimensional regression, we often assume that β satisfies a sparse structure. Let V be the number of false selected features and let R be total number of identified features. The FDP, FDR and k-FWER are respectively defined as

$$\mathrm{FDP} = \frac{V}{\max\{R, 1\}}, \ \ \mathrm{FDR} = \mathbb{E}(\mathrm{FDP})$$

and

$$k\text{-FWER} = \mathrm{Prob}\{V \geq k\}.$$

Moreover, the main notations used in this paper are summarized in *Appendix*. [1].

SLOPE

SLOPE is proposed by Bogdan et al. (2015) for controlled feature selection in high dimensional sparse cases, which replaces the ℓ_1 penalty in Lasso (Tibshirani 1996) with the sorted ℓ_1 penalty. The learning scheme of SLOPE (Bogdan et al. 2015) is formulated as

$$\arg\min_{\beta \in \mathbb{R}^m} \frac{1}{2}||y - X\beta||^2 + \sum_{i=1}^{m} \lambda_i |\beta|_{(i)}, \tag{3}$$

where the regularization parameters $\lambda_1 \geq \cdots \geq \lambda_m \geq 0$ and the regression coefficients $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(m)}$ are all non-negative non-decreasing sequences. When $\lambda_1 = \lambda_2 = \cdots = \lambda_m$, the optimizing scheme (3) obviously reduces to the Lasso (Tibshirani 1996). Given a desired level q, SLOPE controls FDR using the sequence of parameters $\lambda_{\mathrm{BH}} = \{\lambda_{\mathrm{BH}}(1), \lambda_{\mathrm{BH}}(2), \cdots, \lambda_{\mathrm{BH}}(m)\}$ with

$$\lambda_{\mathrm{BH}}(i) = \sigma \cdot \Phi^{-1}(1 - \frac{iq}{2m}), \tag{4}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution under orthogonal design.

Theorem 1 *(Bogdan et al. 2015) In the linear model with the orthogonal design X and $\epsilon \sim N(0, \sigma^2 I_n)$, the SLOPE* (3) *with the regularization parameter sequence (4) satisfies*

$$\mathrm{FDR} \leq q\frac{m_0}{m},$$

where m_0 is the number of true null hypotheses and q is the desired FDR level.

Theorem 1 illustrates the theoretical guarantee of FDR control for SLOPE equipped with λ_{BH} induced by the BH procedure (Benjamini and Hochberg 1995). In this paper, we are not limited to the FDR control, but extend to the k-FWER and FDP control by replacing the BH procedure with the stepdown procedure (Lehmann and Romano 2005).

Algorithm 1: Accelerated proximal gradient algorithm for SLOPE (3)

```
Input: Training set $X \in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^n$ and parameter $\lambda = (\lambda_1, \lambda_2, ..., \lambda_m)$.
Initialization: $a^0 \in \mathbb{R}^m$, $b^0 = a^0$ and $\theta_0 = 1$.
  for k = 0,1,$\cdots$ do
    $b^{k+1} = \mathrm{prox}_{t_k J_\lambda}\left(a^k - t_k X'\left(Xa^k - y\right)\right)$
    $\theta_{k+1}^{-1} = \frac{1}{2}\left(1 + \sqrt{1 + 4/\theta_k^2}\right)$
    $a^{k+1} = b^{k+1} + \theta_{k+1}\left(\theta_k^{-1} - 1\right)\left(b^{k+1} - b^k\right)$
  end for
Output: $a$ satisfying the stopping criteria.
```

From the computing side, the optimization objective function of SLOPE (3) is convex but non-smooth, which can be implemented efficiently by the proximal gradient descent algorithm (Bogdan et al. 2015). For completeness, we state the computing steps of SLOPE in Algorithm 1, which also suits for our variants of SLOPE. Here, $J_\lambda = \sum_{i=1}^{m} \lambda_i |\beta|_{(i)}$ and the step lengths get by backtracking line search and satisfy $t_k < 2/||X||^2$ (Beck and Teboulle 2009; Becker, Candès, and Grant 2011). Moreover, Bogdan et al. (2015) also derive concrete stopping criteria through duality theory.

Stepdown Procedure

The stepdown procedure (Lehmann and Romano 2005) aims to control k-FWER and FDP, i.e., given $\alpha, r \in (0, 1)$,

$$k\text{-FWER} \leq \alpha \tag{5}$$

and

$$\mathrm{Prob}\{\mathrm{FDP} > \gamma\} \leq \alpha. \tag{6}$$

Suppose that there are m individual tests $H_1, ..., H_m$, whose corresponding p-values are $\hat{p}_1, ..., \hat{p}_m$. Let $\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq ... \leq \hat{p}_{(m)}$ be the ordered p-values and let the non-negative non-decreasing sequence $\alpha_1 \leq \alpha_2 ... \leq \alpha_m$ be the k-FWER thresholds. The hypotheses corresponding to the sorted p-values are defined as $H_{(1)}, H_{(2)} ..., H_{(m)}$. Then the stepdown procedure is defined stepwise as follows:

Step 0: Let $i = 0$.

Step 1: If $\hat{p}_{(i+1)} \geq \alpha_{i+1}$, go to step 2. Otherwise, set $i = i + 1$ and repeat *Step* 1.

Step 2: Reject $H_{(j)}$ for $j \leq k$ and accept $H_{(j)}$ for $j > k$.

In other words, if $p_{(1)} > \alpha_1$, no null hypotheses are rejected. Otherwise, if $H_{(1)}, H_{(2)} ..., H_{(r)}$ are rejected, the largest r satisfies

$$p_{(1)} \leq \alpha_1, p_{(2)} \leq \alpha_2, ..., p_{(r)} \leq \alpha_r. \tag{7}$$

Based on the stepdown procedure, Lehmann and Romano (2005) provided two different thresholds to ensure the k-FWER control and the FDP control, respectively.

Theorem 2 *(Lehmann and Romano 2005) For testing $H_i, i = 1, ..., m$, given k and $\alpha \in (0, 1)$, the stepdown procedure described in (7) with*

$$\alpha_i = \begin{cases} \frac{k\alpha}{m}, & i \leq k \\ \frac{k\alpha}{m+k-i}, & i > k \end{cases} \tag{8}$$

[1] See Appendix http://arxiv.org/abs/2302.10610.

controls the k-FWER, that is, (5) holds.

Theorem 3 *(Lehmann and Romano 2005) For testing $H_i, i = 1, ..., m$, given $\alpha, \gamma \in (0, 1)$, if the p-values of false null hypotheses are independent of the true ones, the stepdown procedure described in (7) with*

$$\alpha_i = \frac{(\lfloor \gamma i \rfloor + 1)\alpha}{m + \lfloor \gamma i \rfloor + 1 - i} \tag{9}$$

controls the FDP in the sense of (6).

Theorems 2 and 3 demonstrate that the stepdown procedure enjoys the theoretical guarantees on the k-FWER control and FDP control under ingenious selections of α_i. Indeed, these theoretical properties of stepdown procedure motivate our designs for new SLOPE algorithms.

Methodology

This section injects the stepdown procedure (Lehmann and Romano 2005) into the classical SLOPE (Bogdan et al. 2015) to formulate new stepdown SLOPEs for controlled feature selection to ensure the k-FWER control and the FDP control. Here, we provide the sequences of tuning parameters under the orthogonal design for the k-FWER control and the FDP control, respectively. Furthermore, we present an intuitive theoretical analysis for the selection of regularization parameters in general setting.

Orthogonal Design

It has been illustrated in Bogdan et al. (2015) that the linking between multiple tests and model selection for SLOPE under the orthogonal design. Following this line, we assume that X is an $n \times m$ dimensional orthogonal matrix, i.e, $X'X = I_m$ and $\epsilon \sim N(0, \sigma^2 I_n)$ is an n-dimensional column vector with known variance. Then, the linear regression model

$$y = X\beta + \epsilon$$

is transformed into

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

It is well known that the problem of selecting effective features can be simplified as a multiple hypothesis test problem. Denote m hypotheses as $H_i : \beta_i = 0, 1 \leq i \leq m$. If H_i is rejected, β_i is considered as an effective feature and vice versa. Bogdan et al. (2015) gave the selection mechanism of regularization parameters for SLOPE through the BH procedure (Benjamini and Hochberg 1995) under the orthogonal design. For brevity, we call the proposed methods as k-SLOPE and F-SLOPE with respect to the control of k-FWER and FDP, respectively.

The regularization scheme of k-SLOPE is formulated as

$$\arg\min_{\beta \in \mathbb{R}^m} \frac{1}{2} \|y - X\beta\|_{l_2}^2 + \sigma \cdot \sum_{i=1}^{m} \lambda_{k\text{-FWER}}(i) |\beta|_{(i)}, \tag{10}$$

where

$$\lambda_{k\text{-FWER}}(i) = \begin{cases} \Phi^{-1}(1 - k\alpha/2m), & i \leq k \\ \Phi^{-1}(1 - k\alpha/2(m + k - i)), & i > k. \end{cases} \tag{11}$$

The k-SLOPE equipped with (11) yields the following theoretical property, which has been proved in *Supplementary Appendix*.

Theorem 4 *In the linear model (2) with the orthogonal matrix X and noise $\epsilon \sim N(0, \sigma^2 I_n)$, given k and $\alpha \in (0, 1)$, the k-FWER of the k-SLOPE model (10) satisfies (5).*

Theorem 4 illustrates that k-SLOPE controls the k-FWER under the orthogonal design, which has been proved in *Appendix*. Although the $\lambda_{k\text{-FWER}}(i)$'s are chosen with reference (Lehmann and Romano 2005), (10) is not equivalent to the stepdown procedure described above. We also empirically support this theoretical guarantee by experimental analysis.

Generally, the number of false selected features that people are willing to abide is directly proportional to the number of identified features. Therefore, we may be no longer concerned about k-FWER, but about FDP. Similar to (10), the convex optimization problem of F-SLOPE is formulated as

$$\arg\min_{\beta \in \mathbb{R}^m} \frac{1}{2} \|y - X\beta\|_{l_2}^2 + \sigma \cdot \sum_{i=1}^{m} \lambda_{\text{FDP}}(i) |\beta|_{(i)}, \tag{12}$$

where

$$\lambda_{\text{FDP}}(i) = \Phi^{-1}(1 - \frac{(\lfloor \gamma i \rfloor + 1)\alpha}{2(m + \lfloor \gamma i \rfloor + 1 - i)}).$$

The selection of regularization parameters also produces the following theoretical guarantee.

Theorem 5 *In the linear model (2) with the orthogonal matrix X and noise $\epsilon \sim N(0, \sigma^2 I_n)$, given $\alpha, \gamma \in (0, 1)$, the FDP of the F-SLOPE model (12) satisfies (6).*

Theorem 5 assures the ability of FDP control for F-SLOPE under the orthogonal design setting, which has been established in *Appendix*. The only difference between the F-SLOPE model (12) and the k-SLOPE model (10) is the selection mechanism of the sequence for penalty parameters. The conclusion is also supported by the later orthogonal experiments. Moreover, the optimization algorithm of k-SLOPE and F-SLOPE is the same as that of SLOPE because they are all convex and non-smooth. More optimization details are present in Algorithm 1.

General Setting

Usually, SLOPE is difficult to establish solid theoretical guarantees for the FDR control in non-orthogonal setting (Bogdan et al. 2015). Hence, k-SLOPE and F-SLOPE may also face the degraded performance under such general setting. Fortunately, Bogdan et al. (2015) used their own qualitative insights to make an intuitive adjustment to the regularization parameter sequence and showed the empirical effectiveness. Analogous to SLOPE, we give the regularization parameter forms of k-SLOPE and F-SLOPE through theoretical analysis in general setting.

Assume k-SLOPE and F-SLOPE correctly detect these features and correctly estimate the signs of the regression coefficients. Let X_S and β_S be the subset of variables associated to $\beta_i \neq 0$ and the value of their coefficients, respectively. The nonzero components estimator is approximated by

$$\hat{\beta}_S \approx (X_S' X_S)^{-1}(X_S' y - \lambda_S) = \hat{\beta}_{\text{LSE}} - (X_S' X_S)^{-1} \lambda_S, \tag{13}$$

t	SLOPE			k-SLOPE			F-SLOPE		
	Prob(FDP $> \gamma$)	FDR	Power	Prob(FDP $> \gamma$)	FDR	Power	Prob(FDP $> \gamma$)	FDR	Power
50	0.450	0.094	1.000	0.001	0.006	1.000	0.003	0.007	1.000
100	0.330	0.092	0.995	0.000	0.002	0.998	0.002	0.005	1.000
200	0.140	0.080	0.999	0.001	0.001	1.000	0.000	0.007	1.000
300	0.000	0.070	1.000	0.002	0.001	1.000	0.000	0.005	0.995
400	0.000	0.058	1.000	0.000	0.001	0.995	0.001	0.004	0.994
500	0.000	0.050	1.000	0.000	0.001	0.997	0.000	0.005	0.997

Table 1: Results for controlled feature selection under the orthogonal design (different t and fixed $k = 5$).

where $\lambda_S = (\lambda_1, ..., \lambda_{|S|})'$ and $\hat{\beta}_{\text{LSE}}$ is the least-squares estimator of β_S. Inspired by (Bogdan et al. 2015), we calculate the distribution of $X_i' X_S(\beta_S - \hat{\beta}_S)$ to determine the specific forms of the regularization parameters for k-SLOPE and F-SLOPE. In light of (13),

$$\mathbb{E}(\beta_S - \hat{\beta}_S) \approx (X_S' X_S)^{-1} \lambda_S$$

and

$$\mathbb{E} X_i' X_S(\beta_S - \hat{\beta}_S) \approx \mathbb{E} X_i' X_S (X_S' X_S)^{-1} \lambda_S.$$

Under the gaussian design, where each element of X is i.i.d $N(0, 1/n)$,

$$\begin{aligned}\mathbb{E}(X_i' X_S (X_S' X_S)^{-1} \lambda_S)^2 &= \frac{1}{n} \lambda_S' \mathbb{E}(X_S' X_S)^{-1} \lambda_S \\ &= w(|S|) \cdot ||\lambda_S||^2,\end{aligned}$$

and

$$w(|S|) = \frac{1}{n - |S| - 1},$$

where $|S|$ is the number of elements of S, $i \notin S$ and the second equation relies on the fact that the expected of an inverse $|S| \times |S|$ Wishart matrix with n degrees of freedom is equal to $I_{|S|}/(n - |S| - 1)$ (Nydick 2012).

The k-SLOPE begins with $\lambda_{k\text{G}} = \lambda_{k\text{-FWER}}(1)$. Then, we take into account the slight increase in variance so that

$$\lambda_{k\text{G}}(2) = \lambda_{k\text{-FWER}}(2) \sqrt{1 + w(2) \lambda_{k\text{G}}(1)^2}.$$

Thus, the sequence of $\lambda_{k\text{G}}$ can be expressed as

$$\lambda_{k\text{G}}(i) = \lambda_{k\text{-FWER}}(i) \sqrt{1 + w(i-1) \sum_{j<i} \lambda_{k\text{G}}(i)^2}. \quad (14)$$

The only difference between F-SLOPE and the k-SLOPE is the selection of the coefficient sequence of the penalty term. Similar with (14), F-SLOPE starts with $\lambda_{\text{FG}} = \lambda_{\text{FDP}}(1)$, and then

$$\lambda_{\text{FG}}(i) = \lambda_{\text{FDP}}(i) \sqrt{1 + w(i-1) \sum_{j<i} \lambda_{\text{FG}}(i)^2}. \quad (15)$$

If the coefficient sequence of the penalty term is an incremental sequence, k-SLOPE and F-SLOPE no longer are the convex optimization problems. Denote $k^* := k(n, m, \alpha)$ as the subscript of global minimum, k-SLOPE and F-SLOPE respectively work with

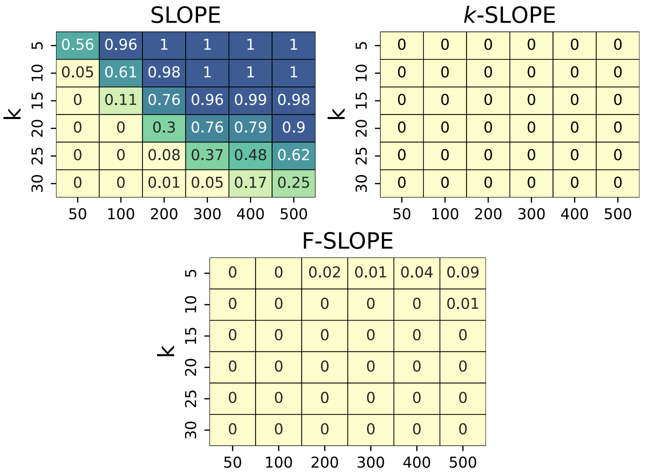

Figure 1: k-FWER provided by different approaches for controlled feature selection under orthogonal design (with different k and t). The value in the small square is the size of k-FWER. The darker the color, the larger the k-FWER and vice versa.

$$\lambda_{\text{kG}^\star}(i) = \begin{cases} \lambda_{\text{kG}}(i), & i \leq k^\star, \\ \lambda_{kG}(k^\star), & i > k^\star, \end{cases} \quad (16)$$

with $\lambda_{k\text{G}}(i)$ given in (14) and

$$\lambda_{\text{FG}}(i) = \begin{cases} \lambda_{\text{FG}}(i), & i \leq k^\star, \\ \lambda_{FG}(k^\star), & i > k^\star, \end{cases} \quad (17)$$

with $\lambda_{\text{FG}}(i)$ defined in (15)). When the design matrix isn't Gaussian or that columns aren't independent, we can employ the Monte Carlo estimate of the correction (Hammersley and Morton 1954) instead of $w(i-1) \sum_{j<i} \lambda(i)^2$ in the formulas (14) and (15).

Empirical Validation

All experiments are implemented in Python on a Macbook Pro with Apple M1 and 16 GB memory. The reported results are the average values after repeating 100 times for each experiment.

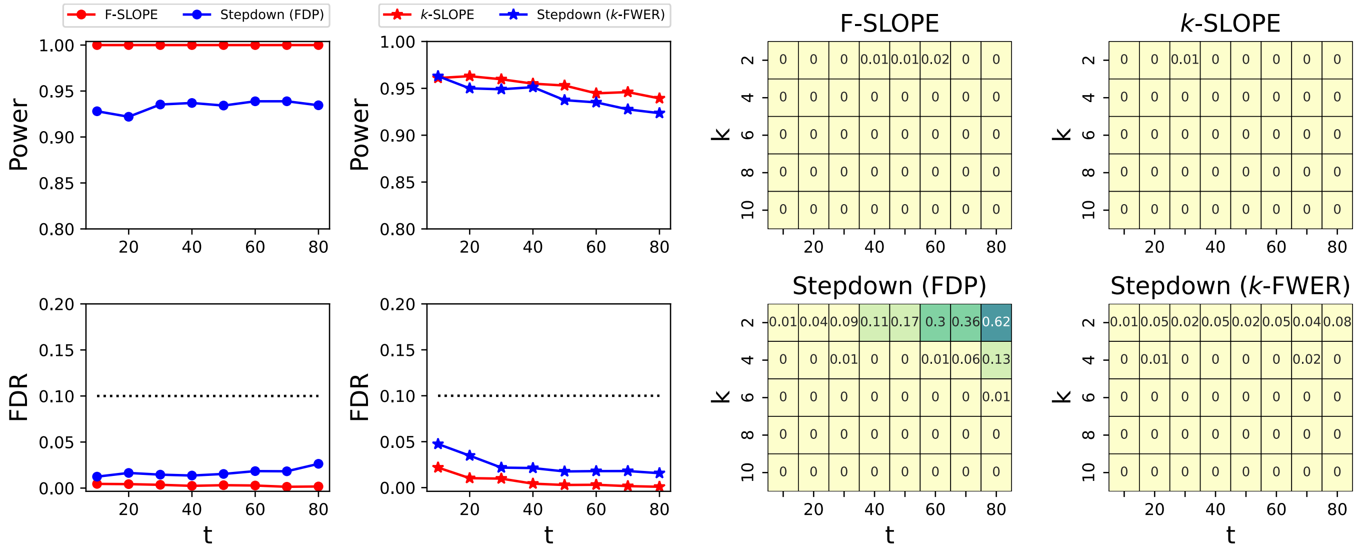

Figure 2: Result for controlled feature selection on the simulated data. The black dashed lines indicate the target FDR level. Constance for k-SLOPE is $k = 6$ in the second column (from left to right). The value in the small square is the size of k-FWER in the third and fourth columns (from left to right). The darker the color, the larger the k-FWER and vice versa.

t	F-SLOPE	k-SLOPE	Sd (FDP)	Sd (k-FWER)
10	0.00	0.04	0.02	0.08
20	0.03	0.00	0.00	0.03
30	0.00	0.01	0.01	0.01
40	0.00	0.00	0.00	0.00
50	0.01	0.00	0.00	0.00
60	0.00	0.00	0.00	0.00
70	0.00	0.00	0.00	0.00
80	0.00	0.00	0.00	0.00

Table 2: $\mathrm{Prob}(\mathrm{FDP} > \gamma)$ results on the simulated data for multiple mean testing ($k = 6$)

Experiments of Orthogonal Design Setting

Inspired by (Bogdan et al. 2015; Brzyski et al. 2019), we draw the design matrix $X = I_n$ with $n = 1000$. Then, we simulate the response from the linear model

$$y = X\beta + \epsilon, \epsilon \sim N(0, I_n).$$

The number of relevant features t is set to vary within $\{50, 100, 200, 300, 400, 500\}$ and the nonzero regression coefficients are equal to $3\sqrt{2\log n}$. We set the target FDR level $\alpha = 0.1$ and $\gamma = 0.1$ for F-SLOPE, and set $k = \{5, 10, 15, 20, 25, 30\}$ and $\alpha = 0.1$ for k-SLOPE. Table 1 reports the estimation of FDR, $\mathrm{Prob}\{\mathrm{FDP} \geq \gamma\}$ and power with 100 repetitions.

Figure 1 summaries the results of SLOPE, k-SLOPE and F-SLOPE in these trials. These results show our proposed stepdown SLOPEs can reach the FDP control, FDR control and k-FWER control flexibly, while SLOPE just can control the FDR. Meanwhile, k-SLOPE and F-SLOPE also enjoy the promising power in almost all settings. Furthermore, these experimental results verify the validity of Theorems 4 and 5. Due to the space limitation, we just present the part experimental results (in Figure 1 and Table 1) and put the comprehensive results in *Appendix*.

Multiple mean testing from correlated statistics

We exemplify the properties of our proposed methods as applied to the typical multiple testing problem with correlated test statistics. Similar to (Bogdan et al. 2015), we consider the following case. Researchers conduct $n = 1000$ experiments in each of $p = 5$ randomly selected laboratories. Observation results are modeled as

$$y_{i,j} = \mu_i + \tau_j + z_{i,j}, \quad 1 \leq i \leq n, 1 \leq j \leq p,$$

where $\tau_j \sim N(0, \sigma_\tau^2)$ is the laboratory impact factors, $z_{i,j} \sim N(0, \sigma_z^2)$ is the errors and they are independent of each other. Our goal is to test whether μ_i is equal to 0, i.e. $H_i : \mu_i = 0, i = 1, 2, ..., n$. Averaging the observed values of 5 laboratories, we get the mean of results

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

where $\bar{y} = (\bar{y}_1, ..., \bar{y}_n)^T$ is drawn independently from $N(\mu, \Sigma)$, where $\Sigma_{i,i} = \frac{1}{5}\sigma_\tau^2 = \rho$ and $\Sigma_{i,j} = \frac{1}{5}(\sigma_\tau^2 + \sigma_z^2) = \sigma^2$ for $i \neq j$ (Bogdan et al. 2015). The key problem is to judge whether the marginal means of a multivariate correlation Gaussian vector disappear or not. One classical solution is to perform marginal tests with $\bar{y}$ statistic, which depends on the stepdown procedure to control k-FWER or FDP (Lehmann and Romano 2005). In other words, we sort the $\bar{y}$ sequence with $|\bar{y}|_{(1)} \geq |\bar{y}|_{(2)} \geq \cdots |\bar{y}|_{(m)}$. Then we use the stepdown procedure with the k-FWER critical values or FDP critical values. Another solution is to "whiten the noise", i.e., the regression equation is reduced to

$$\tilde{y} = \Sigma^{-1/2}\bar{y} = \Sigma^{-1/2}\mu + \epsilon, \tag{18}$$

k / t	weak signals								moderate signals							
	10	20	30	40	50	60	70	80	10	20	30	40	50	60	70	80
2	0.02	0.00	0.02	0.00	0.04	0.04	0.07	0.07	0.00	0.00	0.05	0.03	0.10	0.12	0.07	0.13
4	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
8	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00

Table 3: k-FWER of k-SLOPE ($m = 2n$) on the simulated data under the weak and moderate signals (different t and k).

t	$m = 2n$		$m = n/2$	
	weak	moder	weak	moder
10	0.03	0.00	0.05	0.00
20	0.07	0.00	0.03	0.01
30	0.01	0.00	0.00	0.00
40	0.00	0.01	0.00	0.00
50	0.00	0.00	0.00	0.00
60	0.00	0.00	0.00	0.00
70	0.01	0.00	0.00	0.01
80	0.00	0.02	0.00	0.00

Table 4: Prob(FDP$>\gamma$) of F-SLOPE on the simulated data under the weak and moderate signals (different t).

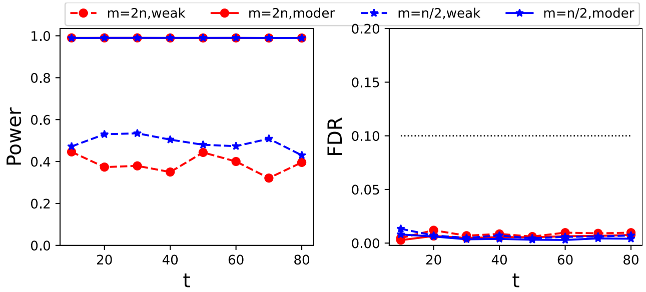

Figure 3: Power and FDR of F-SLOPE under Gaussian design (different t). The black dashed line indicates the target FDR level.

where $\epsilon \sim N(0, I_p)$, $\Sigma^{-1/2}$ is the regression design matrix. If $\Sigma^{-1/2}$ is closed to the orthogonal matrix, the multiple tests problem is transformed into the feature selection problem under the approximate orthogonal design, where k-SLOPE and F-SLOPE can provide better performance.

Similar with (Bogdan et al. 2015), we set $\sigma_\tau^2 = \sigma_z^2 = 2.5$ and consider a sparse setting, where the number of the relevant features t is $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$. The nonzero mean is set to $2\sqrt{2\log p}/c$, where c is equal to the Euclidean norm of each of the columns of $\Sigma^{-1/2}$. We set $\alpha = \gamma = 0.1$ for all FDP controlled methods, and set $k = \{2, 4, 6, 8, 10\}$ and $\alpha = 0.1$ for k-FWER controlled methods. Figure 2 shows the FDR, k-FWER and power provided by F-SLOPE, k-SLOPE, the stepdown procedures for FDP control (Sd(FDP)), and the stepdown procedures for k-FWER control (Sd(k-FWER)). Table 2 shows $\text{Prob}(\text{FDP} > \gamma)$ for F-SLOPE, k-SLOPE and the stepdown procedures. These experimental results show that our proposed methods ensure the FDP, FDR and k-FWER control simultaneously, while Sd (FDP) (or Sd (k-FWER)) focuses on controlling the FDP (or k-FWER) and FDR. However, F-SLOPE and k-SLOPE have greater power than the stepdown procedures. Therefore, our proposed methods have better performance than the classical stepdown procedures in multiple tests. Please refer to *Appendix* for more empirical results.

Experiments of Gaussian Design Setting

We study the performance of k-SLOPE and F-SLOPE in general setting. Following the strategy in (Bogdan et al. 2015), let the entries of the design matrix X are i.i.d $N(0, 1/n)$ with $n = 5000$. The number of relevant features t varies $\{10, 20, 30, 40, 50, 60, 70, 80\}$. Moderate signals having nonzero regression coefficients is set to $2\sqrt{2\log m}$, while this value is set to $\sqrt{2\log m}$ for weak signals. We set $\alpha = \gamma = 0.1$ for F-SLOPE, and set $k = \{2, 4, 6, 8, 10\}$ and $\alpha = 0.1$ for k-SLOPE.

Then we consider two scenarios: (1) $m = 2n$; (2) $m = n/2$. Table 4 and Figure 3 illustrate F-SLOPE keeps the $\text{Prob}(\text{FDP} > \gamma)$ and FDR below the norminal level under both scenarios ($m = 2n$ and $m = n/2$), whether the signals are weak and moderate. Meanwhile, Figure 3 also shows F-SLOPE ($m = n/2$) has greater power than F-SLOPE ($m = 2n$) under weak signals, while F-SLOPE ($m = 2n$) and F-SLOPE ($m = n/2$) have similar power under the moderate signals. As shown in Table 3, k-SLOPE control k-FWER under both scenarios ($m = 2n$ and $m = n/2$). In addition, the power of k-SLOPE also has nice performance under the moderate signals. Moreover, experimental results verify the validity of k-SLOPE with λ_{kG^*} and F-SLOPE with λ_{FG^*}. See *Appendix* for additional experimental results.

Conclusion

This paper formulated two feature selection approaches based on the SLOPE technique (Bogdan et al. 2015). Different from the existing works concerning the FDR control, the current models focus on the k-FWER control and FDP control for feature selection. With the help of stepdown procedure (Lehmann and Romano 2005), we establish their theoretical guarantees under the orthogonal design. Simulated experiments validated the effectiveness of the proposed stepdown SLOPEs and support our theoretical findings.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grant No. 12071166 and by the Fundamental Research Funds for the Central Universities of China under Grant 2662020LXQD002. We are grateful to the anonymous AAAI reviewers for their constructive comments.

References

Aggarwal, S.; and Yadav, A. K. 2016. False discovery rate estimation in proteomics. In *Statistical Analysis in Proteomics*, 119–128. Springer.

Alemán, X.; Duryea, S.; Guerra, N. G.; McEwan, P. J.; Muñoz, R.; Stampini, M.; and Williamson, A. A. 2017. The effects of musical training on child development: A randomized trial of El Sistema in Venezuela. *Prevention Science*, 18(7): 865–878.

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

Barber, R. F.; Candès, E. J.; and Samworth, R. J. 2020. Robust inference with knockoffs. *The Annals of Statistics*, 48(3): 1409–1431.

Beck, A.; and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1): 183–202.

Becker, S. R.; Candès, E. J.; and Grant, M. C. 2011. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical programming computation*, 3(3): 165–218.

Bellec, P. C.; Lecué, G.; and Tsybakov, A. B. 2018. SLOPE meets LASSO: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B): 3603–3642.

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.

Bogdan, M.; Van Den Berg, E.; Sabatti, C.; Su, W.; and Candès, E. J. 2015. SLOPE—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3): 1103.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Brzyski, D.; Gossmann, A.; Su, W.; and Bogdan, M. 2019. Group slope–adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525): 419–433.

Brzyski, D.; Peterson, C. B.; Sobczyk, P.; Candès, E. J.; Bogdan, M.; and Sabatti, C. 2017. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1): 61–75.

Candès, E.; Fan, Y.; Janson, L.; and Lv, J. 2018. Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3): 551–577.

Chen, H.; Wang, X.; Deng, C.; and Huang, H. 2017. Group sparse additive machine. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; and Huang, H. 2021. Sparse modal additive model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6): 2373–2387.

Delattre, S.; and Roquain, E. 2015. New procedures controlling the false discovery proportion via Romano–Wolf's heuristic. *The Annals of Statistics*, 43(3): 1141–1177.

Fan, J.; and Lv, J. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1): 101.

Fan, Y.; Demirkaya, E.; and Lv, J. 2019. Nonuniformity of p-values can occur early in diverging dimensions. *The Journal of Machine Learning Research*, 20(1): 2849–2881.

Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4): 347–388.

Ferreira, J.; and Zwinderman, A. 2006. On the benjamini–hochberg method. *The Annals of Statistics*, 34(4): 1827–1849.

Hammersley, J. M.; and Morton, K. W. 1954. Poor man's monte carlo. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(1): 23–38.

Javanmard, A.; and Javadi, H. 2019. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1): 1212–1253.

Jiang, W.; Bogdan, M.; Josse, J.; Majewski, S.; Miasojedow, B.; Ročková, V.; and Group, T. 2022. Adaptive bayesian SLOPE: Model selection with incomplete data. *Journal of Computational and Graphical Statistics*, 31(1): 113–137.

Kos, M.; and Bogdan, M. 2020. On the asymptotic properties of SLOPE. *Sankhya A*, 82(2): 499–532.

Kremer, P. J.; Lee, S.; Bogdan, M.; and Paterlini, S. 2020. Sparse portfolio selection via the sorted l_1-norm. *Journal of Banking & Finance*, 110: 105687.

Larsson, J.; Bogdan, M.; and Wallin, J. 2020. The strong screening rule for SLOPE. *Advances in Neural Information Processing Systems*, 33: 14592–14603.

Larsson, J.; Klopfenstein, Q.; Massias, M.; and Wallin, J. 2022. Coordinate descent for SLOPE. *arXiv preprint arXiv:2210.14780*.

Lee, S.; Sobczyk, P.; and Bogdan, M. 2019. Structure learning of Gaussian Markov random fields with false discovery rate control. *Symmetry*, 11(10): 1311.

Lehmann, E.; and Romano, J. P. 2005. Generalizations of the familywise error rate. *The Annals of Statistics*, 1138–1154.

Lemhadri, I.; Ruan, F.; and Tibshirani, R. 2021. Lassonet: neural networks with feature sparsity. In *International Conference on Artificial Intelligence and Statistics*, 10–18. PMLR.

Luo, Z.; Sun, D.; Toh, K.-C.; and Xiu, N. 2019. Solving the OSCAR and SLOPE models using a semismooth Newton-based augmented Lagrangian method. *J. Mach. Learn. Res.*, 20(106): 1–25.

Nydick, S. W. 2012. The wishart and inverse wishart distributions. *Electronic Journal of Statistics*, 6(1-19).

Ravikumar, P.; Lafferty, J.; Liu, H.; and Wasserman, L. 2009. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5): 1009–1030.

Riccobello, R.; Bogdan, M.; Bonaccolto, G.; Kremer, P. J.; Paterlini, S.; and Sobczyk, P. 2022. Sparse graphical modelling via the sorted l_1-norm. *arXiv preprint arXiv:2204.10403*.

Romano, J. P.; and Shaikh, A. M. 2006. Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, 34(4): 1850–1873.

Romano, J. P.; and Wolf, M. 2007. Control of generalized error rates in multiple testing. *The Annals of Statistics*, 35(4): 1378–1408.

Romano, Y.; Sesia, M.; and Candès, E. 2020. Deep knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872.

Su, W.; and Candès, E. 2016. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3): 1038–1068.

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

Van der Laan, M. J.; Dudoit, S.; and Pollard, K. S. 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Yu, L.; Kaufmann, T.; and Lederer, J. 2021. False discovery rates in biological networks. In *International Conference on Artificial Intelligence and Statistics*, 163–171. PMLR.

Zhao, X.; Chen, H.; Wang, Y.; Li, W.; Gong, T.; Wang, Y.; and Zheng, F. 2022. Error-based knockoffs inference for controlled feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9190–9198.