Controlling False Discovery Rate Using Gaussian Mirrors

Xin Xing, Zhigen Zhao, Jun S Liu

November 19, 2019

Abstract

Simultaneously finding multiple influential variables and controlling the false discovery rate (FDR) for linear regression models is a fundamental problem with a long history. We here propose the Gaussian Mirror (GM) method, which creates for each predictor variable a pair of mirror variables by adding and subtracting a randomly generated Gaussian random variable, and proceeds with a certain regression method, such as the ordinary least-square or the Lasso. The mirror variables naturally lead to a test statistic highly effective for controlling the FDR. Under a weak dependence assumption, we show that the FDR can be controlled at a user-specified level asymptotically. It is shown that the GM method is more powerful than many existing methods in selecting important variables, subject to the control of FDR especially under the case when high correlations among the covariates exist. The R package is publicly available at https://github.com/BioAlgs/GM.

1 Introduction

Linear regression, which dates back to the beginning of the 19th century, is one of the most important statistical tools for practitioners. The theoretical research addressing various issues arising from big data analyses has gained much attention in the last decade. One important problem is to determine which covariates (aka "predictors") are "useful" or "important" in a linear regression. In early days (before 1970's), people often rely on the t-test to assess the importance of each individual predictor in a regression model, although the method is known to be problematic due to the existence of highly multi-colinearity. A greedy stepwise regression method was later proposed in Efroymson (1960) to alleviate some of the flaws. Good criterion, such as Akaike information criteria (AIC, Akaike (1998)) and Bayesian information criteria (BIC, Schwarz (1978)), for directing its operation were developed later. In recent years, due to the advance of data-generating technologies in both science and industry, researchers discovered that various regularization based regression methods, such as Lasso Tibshirani (1996), SCAD Fan and Li (2001), elastic NET (Zou and Hastie (2005)) and many others, are quite effective in dealing with high-dimensional data and selecting relevant

covariates with certain guaranteed theoretical properties, such as the *selection consistency* and the oracle property.

For the linear regression model, $y = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon$, we are interested in testing p hypotheses $H_j : \beta_j = 0, j = 1, \ldots, p$, simultaneously and finding a statistical rule to decide which hypotheses should be rejected. Let $\mathcal{S}_0 \in \{1, 2, \cdots, p\}$ be the set of variables with $\beta_j = 0$, i.e., not in the true model and let $\mathcal{S}_1 = \mathcal{S}_0^c$. Let $\hat{\mathcal{S}}_1$ be the set that is selected based on a statistical rule. The FDR, quantifying the type I error for this statistical rule, is defined as

$$FDR = \mathbb{E}[FDP], \quad \text{where} \quad FDP = \frac{\#\{i \mid i \in \mathcal{S}_0, i \in \hat{\mathcal{S}}_1\}}{\#\{i \mid i \in \hat{\mathcal{S}}_1\} \vee 1}.$$

The task of controlling FDR at a designated level, say q, is very challenging from two aspects: (i) the test statistic and the corresponding distribution under the null hypothesis is not easily available for high dimensional problems; and (ii) the dependence structure of the covariates induces a complex dependence structure among the estimated regression coefficients. A heuristic method based on permutation tests was introduced by Chung and Romano (2016), but it fails to yield the correct FDR control unless $\mathcal{S}_1$ is empty. Starting with the p-values based on the marginal regression, Fan et al. (2012) proposed a novel way to estimate the false discovery propoption (FDP) when the test statistic follows a multivariate normal distribution. However, the consideration of the marginal regression deviates from the original purpose of the study in many cases.

Barber and Candès (2015) introduced an innovative approach, the *knockoff filter*, which provides a provable FDR control for cases with arbitrary design matrices when $p < n/2$. When $p < n < 2p$, they kept $2p - n$ predictors unchanged and constructed the knockoffs only for the remaining $n - p$ variables. It is guaranteed that the FDR can be controlled with a sacrifice of the power. By introducing knockoffs, they obtained a test statistic that satisfies (i) the symmetric property under the nulls, and (ii) independent signs under the nulls. The key of this method is to construct knockoff variables $\tilde{\boldsymbol{X}}$ such that $\tilde{\boldsymbol{X}}$ preserves the correlation structure of the original $\boldsymbol{X}$. To gain the power, however, one wants to construct knockoffs such that $\tilde{\boldsymbol{X}}$ and $\boldsymbol{X}$ are as dissimilar as possible. When the multi-colinearity between the predictors are high and dense, it leaves little room for one to choose a good knockoff, resulting in a significant decrease in the power. As shown in Section 6, when the correlation between the predictors are high and dense, the knockoffs can still control the FDR, but at a high expense of the power loss.

Later, the knockoff method has been extended to high dimensional (screening + knockoffs Barber and Candès (2019)) and the model-X knockoff approach (Candes et al. (2018)). In the model-X framework, Candes et al. (2018) discarded the linear assumption by considering the joint distribution of the response Y and the predictors X. The goal is to find the Markov blanket, a minimal set $\mathcal{S}$ such that Y is independent of all the others when conditioning on $X_{\mathcal{S}}$. They proposed model-X

knockoffs that can control the FDR. However, this construction relies heavily on the assumption that the distribution of X is completely known, which is generally unrealistic except for some special scenarios. Barber et al. (2018) constructed a robust version of model-X knockoffs based on the estimated joint distribution of $(X_1, \ldots, X_p)$. However, estimating the joint distribution in a high-dimensional setting not only is very challenging even for the multivariate Gaussian case, but also leads to inaccurate FDR controls. For the high dimensional data, an other line of work is the post-selection inference which aims at doing inference conditional on the selection step Berk et al. (2013); Lee et al. (2016). In Lee et al. (2016), the selection event of LASSO is shown to be an union of polyhedra. Tibshirani et al. (2016) provides analogous results for forward stepwise regression and least angle regression Taylor and Tibshirani (2018) extends these results to L_1 penalized likelihood models including generalized regression model.

In this paper, we propose a method called Gaussian Mirror, which constructs the test statistic locally or marginally. Namely, for each variable $\boldsymbol{x}_j$, we create a pair of "mirror variables", $\boldsymbol{x}_j^+ = \boldsymbol{x}_j + c_j\boldsymbol{z}_j$ and $\boldsymbol{x}_j^- = \boldsymbol{x}_j - c_j\boldsymbol{z}_j$, where c_j is a scalar and $\boldsymbol{z}_j \sim N(0, \boldsymbol{I}_n)$. $(\boldsymbol{x}_j^+, \boldsymbol{x}_j^-)$ can be viewed as a special and quantifiable way of perturbing the variable. The perturbation is carefully chosen according to an explicit formula of c_j such that the test statistic we introduce has a symmetric distribution around zero for the null variables. Under the high dimensional case, we proof the symmetric property based on the post-selection theory. Based on this property, we can quantitatively estimate the number of false positives and the FDP. To control the FDR, a data-driven threshold is chosen such that the estimated FDP is no larger than q. By assuming weak dependence conditions on the test statistic, we show that the proposed method controls FDR at q level asymptotically. Adding some perturbations not only helps weaken the dependence among the test statistics, but also leads to conclusions that are stable to perturbations, as advocated in the stability framework of Yu (2013).

The GM procedure calculates the mirror statistics by only perturbing one variable at a time. This simple perturbation enjoys two advantages. First, comparing with the knockoffs and Model-X knockoffs which double the size of the design matrix, we introduce the smallest perturbation to the original design matrix, which can improve the power as shown in our numerical studies. Second, the algorithms of calculating the mirror statistics for $j = 1, \ldots, p$ are completely separable, easily parallelizable, and memory efficient. All of our numerical studies have been implemented using a parallel computation architecture.

The construction of the GM statistics and selection procedure do not require any distributional assumption on the design matrix $\boldsymbol{X}$. In contrast to the model-X knockoffs based methods, the GM design bypasses the difficulty of estimating the joint distribution of $(X_1, \ldots, X_p)$, thus is more flexible in real applications such as GWAS studies where the covariates are types of single nucleotide

polymorphisms (SNPs) taking values in $\{0, 1, 2\}$.

In addition, practitioners often have a limited budget to verify the discoveries. For example, if the budget only allows the scientist to do 100 validation tests, a natural question is how many false discoveries (FDs) they expect to have in a top-100 list and whether statisticians can provide an uncertainty measure for the estimated FDs. To address this practical problem, we provide an estimate of the number of FDs in a given list based on the proposed mirror statistics and use the nonparametric bootstrap method to give a confidence interval for the proposed estimate.

The paper is organized as follows. In Section 2, we propose the general GM idea and a GM algorithm for the ordinary least squares (OLS) estimation in low-dimensional cases ($p < n$). In Section 3, we employ the post-selection adjustment strategy and extend the GM construction for Lasso, which handles cases with $p \geq n$ at ease. In Section 4, we develop theoretical properties of the GM procedure. In Section 5, we introduce an estimator of the number of false discoveries and build its confidence interval using non-parametric bootstrap. In Section 6, we provide numerical evidences showing the advantage of GM to its competitors via simulations and real data analysis.

Notation: To ease the readability, we summarize some notations used frequently in this paper. Let $\boldsymbol{X}$ be the design matrix. Without loss of generality, we assume that $\boldsymbol{X}$ is normalized so that $\frac{1}{n}||\boldsymbol{x}_j||_2 = 1$ for $j = 1, \ldots, n$. Let $\boldsymbol{x}_j$ be the j-th column of $\boldsymbol{X}$ and let $\boldsymbol{X}_{-j}$ be the submatrix of $\boldsymbol{X}$ with the j-th column removed. Denote $\boldsymbol{X}^j = (\boldsymbol{x}_j^+, \boldsymbol{x}_j^-, \boldsymbol{X}_{-j})$, where $\boldsymbol{x}_j^+ = \boldsymbol{x}_j + c_j\boldsymbol{z}_j$, $\boldsymbol{x}_j^- = \boldsymbol{x}_j - c_j\boldsymbol{z}_j$, $\boldsymbol{z}_j \sim N(0, \boldsymbol{I}_n)$, and c_j is a scalar. Let $\boldsymbol{\beta}$ be the vector coefficient and let $\boldsymbol{\beta}_{-j}$ be the subvector with the j-th entry removed. Let $\boldsymbol{\beta}^j = (\beta_j^+, \beta_j^-, \boldsymbol{\beta}_{-j}^\top)^\top$, where β_j^+ and β_j^- denote the coefficients of the mirror variable pair, and let $\hat{\boldsymbol{\beta}}^j$ be the corresponding estimator. Let $\mathcal{S}_0 = \{j : \beta_j = 0\}$ and $\mathcal{S}_1 = \{j : \beta_j \neq 0\}$. Let q be the designated FDR level to be controlled.

2 The Gaussian Mirror Methodology for OLS

2.1 The Gaussian mirror

Suppose the data are generated from the following linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of i.i.d. Gaussian white noise with variance σ^2. We begin with the low dimensional case with $n < p$ and focus on the OLS estimator. The high-dimensional setting with $p > n$ is deferred to the next section. As shown in Figure 1, we construct the j-th *Gaussian Mirror* by replacing $\boldsymbol{x}_j$ with a pair of variables $(\boldsymbol{x}_j^+, \boldsymbol{x}_j^-)$ with $\boldsymbol{x}_j^+ = \boldsymbol{x}_j + c_j\boldsymbol{z}_j$ and $\boldsymbol{x}_j^- = \boldsymbol{x}_j - c_j\boldsymbol{z}_j$, where

$\boldsymbol{z}_j$ is an independently simulated Gaussian random vector with mean 0 and covariance $\boldsymbol{I}_n$ and c_j is a scalar which depends on $\boldsymbol{X}$ and $\boldsymbol{z}$.

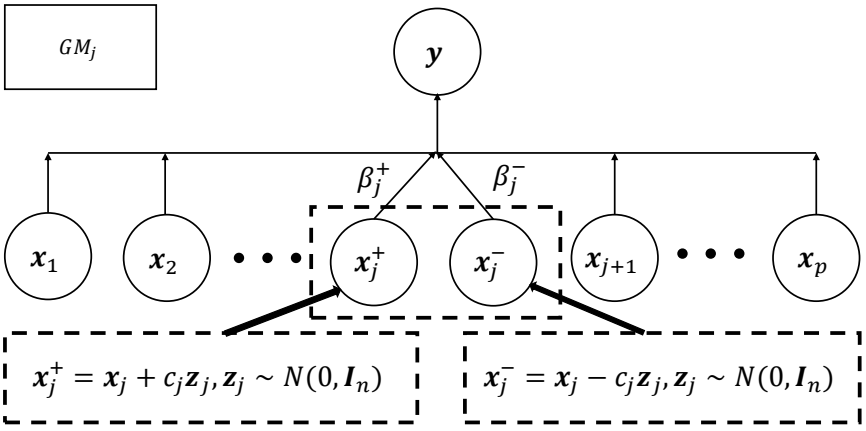

Figure 1: Flowchart of the j-th *Gaussian Mirror*.

Clearly, at the population level we have $\beta_j^+ = \beta_j^- = 0$ if $j \in S_0$, i.e., for variables in the null set; and $\beta_j^+ = \beta_j^- = \beta_j/2$ if $j \in S_1$. Let GM_j be the design matrix with the j-th Gaussian mirror, i.e.,

$$GM_j := \boldsymbol{X}^j = (\boldsymbol{x}_j^+, \boldsymbol{x}_j^-, \boldsymbol{X}_{-j}) = (\boldsymbol{x}_j + c_j\boldsymbol{z}_j, \boldsymbol{x}_j - c_j\boldsymbol{z}_j, \boldsymbol{X}_{-j}). \tag{2}$$

Then we rewrite model in (1) as $\boldsymbol{y} = \boldsymbol{X}^j\boldsymbol{\beta}^j + \boldsymbol{\epsilon}$. We estimate the regression coefficients by the following least squares,

$$\hat{\boldsymbol{\beta}}^j = \underset{\boldsymbol{\beta}^j=(\beta_j^+,\beta_j^-,\boldsymbol{\beta}_{-j})}{\arg\min} ||\boldsymbol{y} - \boldsymbol{X}^j\boldsymbol{\beta}^j||_2^2. \tag{3}$$

We have $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ are both unbiased estimates. We construct the mirror statistics for the j-th variable as

$$M_j = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|. \tag{4}$$

The mirror statistics has two parts: the first part reflects the importance of the j-th predictor; and the second part captures the noise cancellation effect. Intuitively, for a relevant variable $j \in \mathcal{S}_1$, the coefficients $\hat{\beta}_j^+$(or $\tilde{\beta}_j^+$) and $\hat{\beta}_j^-$(or $\tilde{\beta}_j^-$) tend to be close to each other so as to cancel out the noise effect, which results in a relatively large value of M_j. Based on this observation, we regard variable $\boldsymbol{x}_j$ as "significant" when $M_j > t$ for some threshold $t > 0$. If $j \in \mathcal{S}_0$, we choose c_j appropriately such that M_j follows a symmetric distribution with respect to zero, i.e., $P(M_j > s) = P(M_j < -s)$, $\forall s$. Consequently, the number of false positive $\#\{j \in \mathcal{S}_0 \,|\, M_j \geq t\}$ approximately equals $\#\{j \in \mathcal{S}_0 \,|\, M_j \leq -t\}$.

In practice, we do not know the underlying active set $\mathcal{S}_1$. Thus, we use $\#\{j \,|\, M_j \leq -t\}$ as an estimate of the number of false positives, which can be shown to be a slightly over-estimation under

certain conditions, and use

$$\widehat{\mathrm{FDP}}(t) \triangleq \frac{\#\{j \,|\, M_j \leq -t\}}{\#\{j \,|\, M_j \geq t\} \vee 1} \tag{5}$$

as an estimate of the FDP. For any designated FDR level q, $\widehat{\mathrm{FDP}}$ leads to a natural choice of the cutoff τ_q such that

$$\tau_q = \min_t \{\widehat{\mathrm{FDP}}(t) \leq q\}. \tag{6}$$

Based on the data-driven threshold in (6), we select the set of variables as $\hat{\mathcal{S}}_1 = \{j \,|\, M_j \geq \tau_q\}$.

The key to achieve a reasonable estimate of the FDP is to construct the Gaussian mirror GM_j to guarantee the symmetric property of M_j when $j \in \mathcal{S}_0$, which, as shown in the next subsection, can be achieved by carefully choosing the scalar c_j. We discuss GM constructions for the OLS estimates (requiring $n > p$), and then extend the results to the Lasso estimates for high-dimensional cases.

2.2 Gaussian mirrors for the OLS estimator

The j-th GM design ($j = 1, \cdots, p$) for the OLS estimates lead to the following quantity:

$$\hat{\boldsymbol{\beta}}^j := (\hat{\beta}_j^+, \hat{\beta}_j^-, \hat{\boldsymbol{\beta}}_{-j}) = \underset{\boldsymbol{\beta}^j = (\beta_j^+, \beta_j^-, \boldsymbol{\beta}_{-j})}{\arg\min} ||\boldsymbol{y} - \boldsymbol{X}^j \boldsymbol{\beta}^j||_2^2, \tag{7}$$

which has an explicit expression as $\hat{\boldsymbol{\beta}}^j = ((\boldsymbol{X}^j)^\top \boldsymbol{X}^j)^{-1} (\boldsymbol{X}^j)^\top \boldsymbol{y}$. It is known that in $\hat{\boldsymbol{\beta}}^j$, $(\hat{\beta}_j^+, \hat{\beta}_j^-)$ follow a bi-variate normal distribution with mean zero conditional on $\boldsymbol{X}^j$. The following result provides a sufficient condition for the mirror statistics $M_j = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|$ being symmetrically distributed with $j \in \mathcal{S}_0$.

Proposition 1. *Suppose U and V are two random variables following a bivariate normal distribution with mean zero. If the correlation between U and V is zero, we have $M = |U + V| - |U - V|$ following a symmetric distribution about zero, i.e., $P(M > t) = P(M < -t)$, $\forall t > 0$.*

Proof. Since the correlation between U and V is zero, we have $\mathrm{Var}(U + V) = \mathrm{Var}(U - V)$. Furthermore, combining with the fact that $\mathbb{E}(U + V) = \mathbb{E}(U - V) = 0$, the joint distribution of $(U + V, U - V)$ is identical to $(U - V, U + V)$. Thus, M and $-M$ follow the same distribution, i.e., $P(M > t) = P(M < -t)$, $\forall t > 0$. □

The following lemma provides an explicit construction of c_j such that the correlation between $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ is zero, resulting in a symmetric distribution of M_j.

Lemma 1. *For the GM_j design in (2), we can choose*

$$c_j = \sqrt{\frac{\boldsymbol{x}_j^\top (\boldsymbol{I}_n - \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1} \boldsymbol{X}_{-j}^\top) \boldsymbol{x}_j}{\boldsymbol{z}_j^\top (\boldsymbol{I}_n - \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top \boldsymbol{X}_{-j})^{-1} \boldsymbol{X}_{-j}^\top) \boldsymbol{z}_j}}, \tag{8}$$

so that the correlation between $\hat{\beta}_j^+$ *and* $\hat{\beta}_j^-$ *given* $\boldsymbol{z}_j$ *and* $\boldsymbol{X}_{-j}$ *is zero.*

Proof. Conditioning on $\boldsymbol{X}_j$, the covariance matrix of $\hat{\boldsymbol{\beta}}^j$ is

$$\frac{1}{n}\mathrm{Cov}(\hat{\boldsymbol{\beta}}^j \mid \boldsymbol{X}^j) = \sigma^2 \begin{bmatrix} 1+v^{\boldsymbol{z}_j}+2\rho^{(\boldsymbol{x}_j,\boldsymbol{z}_j)} & 1-v^{\boldsymbol{z}_j} & (\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}+\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top \\ 1-v^{\boldsymbol{z}_j} & 1+v^{\boldsymbol{z}_j}-2\rho^{(\boldsymbol{x}_j,\boldsymbol{z}_j)} & (\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}-\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top \\ \rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}+\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})} & \rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}-\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})} & \Sigma_{-j} \end{bmatrix}^{-1}. \tag{9}$$

where $v^{\boldsymbol{z}_j} = \frac{1}{n}c_j^2\boldsymbol{z}_j^\top\boldsymbol{z}_j$, $\rho^{(\boldsymbol{x}_j,\boldsymbol{z}_j)} = \frac{1}{n}c_j\boldsymbol{x}_j^\top\boldsymbol{z}_j$, $(\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})})^\top = \frac{1}{n}\boldsymbol{x}_j^\top\boldsymbol{X}_{-j}$, $(\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top = \frac{1}{n}c_j\boldsymbol{z}_j^\top\boldsymbol{X}_{-j}$, and $\Sigma_{-j} = \frac{1}{n}\boldsymbol{X}_{-j}^\top\boldsymbol{X}_{-j}$. Through calculating the inverse of the block matrix in (9), we have the covariance of $(\hat{\beta}_j^+, \hat{\beta}_j^-)$ given $\boldsymbol{X}^j$ as

$$\begin{aligned} &\frac{1}{n}\mathrm{Cov}(\hat{\beta}_j^+, \hat{\beta}_j^- \mid \boldsymbol{X}^j) \\ =&K^{-1}\sigma^2\left(1-v^{\boldsymbol{z}_j}-(\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}+(\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})}\right), \end{aligned} \tag{10}$$

where

$$\begin{aligned} K^{-1} \;=\; & (1+v^{\boldsymbol{z}_j})^2-4(\rho^{(\boldsymbol{x}_j,\boldsymbol{z}_j)})^2-((\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}+(\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^2 \\ & -(1-v^{\boldsymbol{z}_j}-(\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}+(\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^2. \end{aligned}$$

Simple calculation shows that if we choose c_j such that $v^{\boldsymbol{z}_j}-(\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{z}_j,\boldsymbol{X}_{-j})} = 1-(\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})})^\top\Sigma_{-j}^{-1}\rho^{(\boldsymbol{x}_j,\boldsymbol{X}_{-j})}$, which is equivalant to

$$c_j^2\boldsymbol{z}_j^\top(\boldsymbol{I}_n-\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^\top\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top)\boldsymbol{z}_j = \boldsymbol{x}_j^\top(\boldsymbol{I}_n-\boldsymbol{X}_{-j}(X_{-j}^\top\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^\top)\boldsymbol{x}_j,$$

then the covariance between $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ is zero. We note that the numerator and denominator of c_j in (8) are the lengths of the projections of $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ onto the space of $\boldsymbol{X}_{-j}$'s orthogonal complement, respectively. □

Consequently, to construct the GM_j for $j = 1, \cdots, p$, we first generate the random vector $\boldsymbol{z}_j$ from $N(0, \boldsymbol{I}_n)$, and then choose c_j as (8) such that the covariance between $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ is zero. We summarize the construction of the Gaussian mirror as follows.

Definition 1. *(Gaussian Mirror for OLS) For* $j = 1, \cdots, p$*, one first generates* n*-dimensional Gaussian random vectors* $\boldsymbol{z}_j$ *from* $N(0, \boldsymbol{I}_n)$*, and then computes* c_j *based on equation (8). The* j*-th GM is designed as* $GM_j = \{\boldsymbol{x}_j + c_j\boldsymbol{z}_j, \boldsymbol{x}_j - c_j\boldsymbol{z}_j, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p\}$.

As we have argued, the GM_j defined in Definition 1 guarantees that the covariance between $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ is zero based on Lemma 1. We further construct the mirror test statistics M_j as in (4). The following theorem is a direct consequence of Proposition 1.

Theorem 1. *Let M_j be the test statistics defined in (4) based on GM_j in Definition 1, then*

$$P(M_j < -t \mid \boldsymbol{z}_j) = P(M_j > t \mid \boldsymbol{z}_j),\ \forall t > 0,$$

for $j \in \mathcal{S}_0$.

Algorithm 1 Gaussian mirror algorithm for OLS

1. Parallel FOR $j = 1, 2, \cdots, p$:

   (a) Generate $\boldsymbol{z}_j$ from Gaussian distribution with mean 0 and variance $\boldsymbol{I}_n$.

   (b) Calculate c_j using (8) and get the j-th GM design, GM_j, as in Definition 1.

   (c) Obtain the ordinary least square estimator of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$:

   $$(\hat{\beta}_j^+, \hat{\beta}_j^-, \hat{\boldsymbol{\beta}}_{-j}) = \underset{\beta_j^+, \beta_j^-, \boldsymbol{\beta}_{-j}}{\arg\min}\ ||\boldsymbol{y} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{x}_j^+\beta_j^+ - \boldsymbol{x}_j^-\beta_j^-||_2^2.$$

   (d) Calculate the mirror statistics $M_j = |\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|$.

   END parallel FOR loop

2. For a designated FDR level q, calculate the cutoff τ_q as

   $$\tau_q = \min_t \left\{ t : \frac{\#\{j \mid M_j \le -t\}}{\#\{j \mid M_j \ge t\} \vee 1} \le q \right\}.$$

3. Output the index of the selected variables: $\hat{\mathcal{S}}_1 = \{j \mid M_j \ge \tau_q\}$.

The GM design is a special type of data perturbation method. Perturbation has been widely used in statistics to ensure stability and quantify uncertainty (Yu and Kumbier, 2019; Yu, 2013). For example, jackknife and bootstrap (Efron, 1992) are efficient data perturbation methods with wide applications in statistical inference. In this work, we aim to control FDR based on a new type of data perturbation. The perturbation in the GM design can reduce correlations between the mirrored predictor and other ones, likely improving the robustness and power of variable selection. Theorem 1 guarantees that the distribution of M_j is symmetric with respect to zero for the null variables. Therefore, for any threshold t, if we select the variables with the mirror statistics $M_j > t$, a natural estimate of the FDP is given by (5). For a designated FDR level q, a data-driven threshold τ_q can be obtained as in (6). Note that the GM construction enables independent calculations of the M_j's for $j = 1, \ldots, p$, and is easily parallelizable. Algorithm 1 can thus be implemented on a parallel computing architecture, which significantly reduces the computational time. In Section 4,

we will show that the data-driven choice of τ_q in Algorithm 1 guarantees that FDR is controlled asymptotically under some weak dependence assumptions of the mirror statistics.

3 Gaussian Mirrors for High Dimensional Regression

In high-dimensional settings with the number of features p greater than the number of observations n, one can still create mirror variables $\boldsymbol{x}_j^+ = \boldsymbol{x}_j + c_j \boldsymbol{z}_j$ and $\boldsymbol{x}_j^- = \boldsymbol{x}_j - c_j \boldsymbol{z}_j$, for $j = 1, \cdots, p$, and construct the mirror statistics naturally using the Lasso estimator in (14). Although this simple extension appears to work quite well empirically, its theoretical justification is challenging due to the following reasons: (i) The specific design of c_j as in the OLS case is no longer available. (ii) The Lasso estimator is biased because of the regularization resulting from L_1 penalty with $\lambda_n > 0$. This implies that $\mathbb{E}(\hat{\beta}_j^+) \neq 0$ and $\mathbb{E}(\hat{\beta}_j^-) \neq 0$ for j in the false discovered set $\{j \in \mathcal{S}_0, \hat{\beta}_j^+ \neq 0, \hat{\beta}_j^- \neq 0\}$. (iii) The L_1 penalization also forces a linear constraint on $\mathbf{y}$ in the selection procedure. Since the Lasso estimator is a nonlinear function of $\mathbf{y}$, the constraint on the Lasso estimator is nonlinear, leading to its complicated distribution. We here propose a post-selection strategy to design the GM and construct the mirror statistics M_j such that each M_j satisfies the symmetric property when $j \in \mathcal{S}_0$. Then, a consistent estimate of FDR can be derived via (5).

3.1 Literature on high-dimensional inference

To enable a proper statistical inference in high dimensional settings involving variable selections, Zhang and Zhang (2014) and Van de Geer et al. (2014) propose de-sparsified Lasso; Wasserman and Roeder (2009) advocate using data splitting (Cox, 1975) to avoid dealing with complex constraints induced by variable selection; and Berk et al. (2013) suggest the post-selection adjustments framework, aiming to provide valid statistical inference conditioning on the data-driven variable selection result. In data splitting, the data is divided into two parts, with one half used to select the model and the other half to conduct inference. For post-selection adjustments, (Berk et al., 2013), Lee et al. (2016) propose a procedure that first selects variables using Lasso and then obtains the OLS estimates for the selected model. By characterizing the selection event as a series of linear constraints on the post-selection OLS estimates, they provide valid post-selection inference on certain linear combinations of the coefficients of the selected variables.

Our goal is to control the number of false selected variables based on mirror statistics, which requires that the mirror statistics be symmetrically distributed for null variables. To characterize the distribution of mirror statistics on the null variable set, we consider the post-selection procedure based on Lasso. More specifically, similar to Lee et al. (2016), we first use Lasso to select the model and then re-fit the model with OLS and construct mirror statistics the same way as in the

low-dimensional case. In the following, we explain how to adjust such constructed mirror statistics to accommodate the fact that the fitted model has to be conditional on the selection event.

3.2 Post-selection adjustment for Gaussian mirrors

Recall that Lasso solves the minimization problem

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_n ||\boldsymbol{\beta}||_1. \tag{11}$$

Denote $\hat{\mathcal{S}} = \{j : \tilde{\beta}_j \neq 0\}$ and $\hat{\mathbf{s}} = \text{sign}(\tilde{\beta}_{\hat{\mathcal{S}}})$ as the active set and the related sign based on Lasso estimator. By Lemma 4.1 in Lee et al. (2016), the event $\{(\hat{\mathcal{S}}, \hat{\mathbf{s}}) = (\mathcal{S}, \mathbf{s})\}$ can be rewritten as a series of constrains on $\mathbf{y}$ as follows

$$\{\hat{\mathcal{S}} = \mathcal{S}, \hat{\mathbf{s}} = \mathbf{s}\} = \left\{ \begin{pmatrix} A_0(\mathcal{S}, \mathbf{s}) \\ A_1(\mathcal{S}, \mathbf{s}) \end{pmatrix} \mathbf{y} \leq \begin{pmatrix} \boldsymbol{b}_0(\mathcal{S}, \mathbf{s}) \\ \boldsymbol{b}_1(\mathcal{S}, \mathbf{s}) \end{pmatrix} \right\}, \tag{12}$$

where A_0 encodes the "inactive" constraints determine the selection set, and A_1 encodes the "active" constraints which determine the sign of nonzero coefficients. The expression of these matrices are

$$\begin{aligned} A_0(\mathcal{S}, \mathbf{s}) &= \frac{1}{\lambda} \begin{pmatrix} X_{-\mathcal{S}}^\top (I - P_{\mathcal{S}}) \\ -X_{-\mathcal{S}}^\top (I - P_{\mathcal{S}}) \end{pmatrix} \\ \boldsymbol{b}_0(\mathcal{S}, \mathbf{s}) &= \begin{pmatrix} 1 - X_{-\mathcal{S}}^\top (X_{\mathcal{S}})^+ \mathbf{s} \\ 1 + X_{-\mathcal{S}}^\top (X_{\mathcal{S}})^+ \mathbf{s} \end{pmatrix} \\ A_1(\mathcal{S}, \mathbf{s}) &= -\text{diag}(\mathbf{s})(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} X_{\mathcal{S}}^\top \\ \boldsymbol{b}_1(\mathcal{S}, \mathbf{s}) &= -\lambda \text{diag}(\mathbf{s})(X_{\mathcal{S}}^\top X_{\mathcal{S}})^{-1} \mathbf{s}, \end{aligned} \tag{13}$$

where $P_{\mathcal{S}}$ is the projection matrix of $\mathcal{S}$.

We now develop the mirror statistics conditional on the event $\{(\hat{\mathcal{S}}, \hat{\mathbf{s}}) = (\mathcal{S}, \mathbf{s})\}$, and focus on the design matrix $\boldsymbol{X}_{\mathcal{S}}$. We first define the Gaussian mirrors for the j-th variable in $\mathcal{S}$, $j = 1, \ldots, |\mathcal{S}|$. Denote $\boldsymbol{X}_{-j(\mathcal{S})}$ as the the submatrix of $\boldsymbol{X}_{\mathcal{S}}$ by removing its j-th column. Then we choose c_j as

$$c_j = \sqrt{\frac{\boldsymbol{x}_j^\top (\boldsymbol{I}_n - \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j(\mathcal{S})}^\top \boldsymbol{X}_{-j(\mathcal{S})})^{-1} \boldsymbol{X}_{-j(\mathcal{S})}^\top) \boldsymbol{x}_j}{\tilde{\boldsymbol{z}}_j^\top (\boldsymbol{I}_n - \boldsymbol{X}_{-j(\mathcal{S})}(\boldsymbol{X}_{-j(\mathcal{S})}^\top \boldsymbol{X}_{-j(\mathcal{S})})^{-1} \boldsymbol{X}_{-j(\mathcal{S})}^\top) \tilde{\boldsymbol{z}}_j}}, \tag{14}$$

where $\tilde{\boldsymbol{z}}_j = (I - P_{\mathcal{S}})\boldsymbol{z}_j$ with $\boldsymbol{z}_j$ generated from $N(0, \boldsymbol{I}_n)$. Comparing with (8) in the OLS case, c_j is constructed by first projecting $\boldsymbol{z}_j$ on the orthognal space of $\boldsymbol{X}_{\mathcal{S}}$. Based on c_j, we define the Gaussian mirror designs as follows.

Definition 2. *(Post-selection Gaussian Mirror) Given the post-selection set $\mathcal{S}$ and the corresponding design matrix $\boldsymbol{X}_{\mathcal{S}}$, for $j \in \{1, \ldots, |\mathcal{S}|\}$, we generate n-dimensional random vector $\boldsymbol{z}_j$ from $N(0, \boldsymbol{I}_n)$,*

then compute c_j *based on equation (14). The* j*-th GM is designed as* $\boldsymbol{X}^j_{\mathcal{S}} = \{\boldsymbol{x}_j + c_j\boldsymbol{z}_j, \boldsymbol{x}_j - c_j\boldsymbol{z}_j, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_{|\mathcal{S}|}\}$.

Let $\boldsymbol{\eta}_1^\top = \mathbf{e}_1^\top((\boldsymbol{X}^j_{\mathcal{S}})^\top\boldsymbol{X}^j_{\mathcal{S}})^{-1}(\boldsymbol{X}^j_{\mathcal{S}})^\top$ and $\boldsymbol{\eta}_2^\top = \mathbf{e}_2^\top((\boldsymbol{X}^j_{\mathcal{S}})^\top\boldsymbol{X}^j_{\mathcal{S}})^{-1}(\boldsymbol{X}^j_{\mathcal{S}})^\top$ where $\mathbf{e}_\ell$ is the standard basis vector with the ℓth entry as one and the others as zero for $\ell = 1, 2$. $\boldsymbol{\eta}_1^\top\mathbf{y}$ and $\boldsymbol{\eta}_2^\top\mathbf{y}$ are the first two dimensional OLS estimates of y regress on $\boldsymbol{X}^j_{\mathcal{S}}$. That is,

$$\hat{\beta}_j^+ = \boldsymbol{\eta}_1^\top\mathbf{y} \text{ and } \hat{\beta}_j^- = \boldsymbol{\eta}_2^\top\mathbf{y}. \tag{15}$$

Before deriving the joint post-selection distribution of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$, we first characterize the linear constrains on $(\hat{\beta}_j^+ + \hat{\beta}_j^-, \hat{\beta}_j^+ - \hat{\beta}_j^-)$ resulted from the post-selection event $\mathcal{S}$.

Lemma 2. *Let* $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$*, we have*

$$(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)^\top(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) = 0, \quad \boldsymbol{\eta}_1^T\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2^T\boldsymbol{\eta}_2 = \alpha. \tag{16}$$

Let $A_0(\mathcal{S}, \mathbf{s})$ *and* $A_1(\mathcal{S}, \mathbf{s})$ *be matrices defined in (13), then there exist* $\mathbf{a}_0 \in \mathbb{R}^{2(p-|\mathcal{S}|)}$ *and* $\mathbf{a}_1 \in \mathbb{R}^{|\mathcal{S}|}$*, such that*

$$A_0(\mathcal{S}, \mathbf{s})\boldsymbol{\eta} = \mathbf{a}_0(1, -1) \text{ and } A_1(\mathcal{S}, \mathbf{s})\boldsymbol{\eta} = \mathbf{a}_1(1, 1). \tag{17}$$

In Lemma 2 equation (17), we show that $A_0(\mathcal{S}, \mathbf{s})\boldsymbol{\eta}$ and $A_1(\mathcal{S}, \mathbf{s})\boldsymbol{\eta}$ are both rank one. Write $\mathbf{y} = P_{\boldsymbol{\eta}}\mathbf{y} + (I - P_{\boldsymbol{\eta}})\mathbf{y}$, and let $\mathbf{r} = (I - P_{\boldsymbol{\eta}})\mathbf{y}$ and it is easy to verify that $\mathbf{r}$ is uncorrelated with $\boldsymbol{\eta}^\top\mathbf{y}$, hence independent of $\boldsymbol{\eta}^\top\mathbf{y}$. Then the constrain $A_0(\mathcal{S}, \mathbf{s})\mathbf{y} \le \boldsymbol{b}_0(\mathcal{S}, \mathbf{s})$ in (13) is equivalent to

$$\begin{aligned} A_0(\mathcal{S}, \mathbf{s})\boldsymbol{\eta}(\boldsymbol{\eta}^\top\boldsymbol{\eta})^{-1}\boldsymbol{\eta}^\top\mathbf{y} + A_0(\mathcal{S}, \mathbf{s})\mathbf{r} &= \mathbf{a}_0(1, -1)\text{diag}(\alpha, \alpha)(\hat{\beta}_j^+, \hat{\beta}_j^-)^\top + A_0(\mathcal{S}, \mathbf{s})\mathbf{r} \\ &= \alpha\mathbf{a}_0(\hat{\beta}_j^+ - \hat{\beta}_j^-) + A_0(\mathcal{S}, \mathbf{s})\mathbf{r} < \boldsymbol{b}_0(\mathcal{S}, \mathbf{s}); \end{aligned}$$

i.e., the "inactive" constraints on $(\hat{\beta}_j^+, \hat{\beta}_j^-)$ are applied on the direction of $[1, -1]$. Similarly, we have $\alpha\mathbf{a}_1(\hat{\beta}_j^+ + \hat{\beta}_j^-) + A_1(\mathcal{S}, \mathbf{s})\mathbf{r} < \boldsymbol{b}_1(\mathcal{S}, \mathbf{s}))$, that is, the "active" constraints are applied on the direction of $[1, 1]$. As shown in Figure 2 (a), the constraint regions for $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$ are along the line with slope 1 and -1. By rotating the coordinate system by 45°, we have the joint distribution of $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$ shown in Figure 2 (b), where the constraint regions are parallel to the x-axis and y-axis. We characterize the constraints provided by the selection event $\mathcal{S}$ in the following Lemma 3.

Lemma 3. *The selection event can be written as follows:*

$$\begin{aligned} &\{A\mathbf{y} \le \boldsymbol{b}\} := \{A_0(\mathcal{S}, \mathbf{s})\mathbf{y} \le \boldsymbol{b}_0(\mathcal{S}, \mathbf{s})\} \cap \{A_1(\mathcal{S}, \mathbf{s})\mathbf{y} \le \boldsymbol{b}_1(\mathcal{S}, \mathbf{s})\} \\ =&\{\mathcal{V}_1^L(\mathbf{r}) \le \hat{\beta}_j^+ + \hat{\beta}_j^- \le \mathcal{V}_1^U(\mathbf{r}),\ \mathcal{V}_1^N(\mathbf{r}) > 0\} \cap \{\mathcal{V}_0^L(\mathbf{r}) \le \hat{\beta}_j^+ - \hat{\beta}_j^- \le \mathcal{V}_0^U(\mathbf{r}),\ \mathcal{V}_0^N(\mathbf{r}) > 0\} \end{aligned}$$

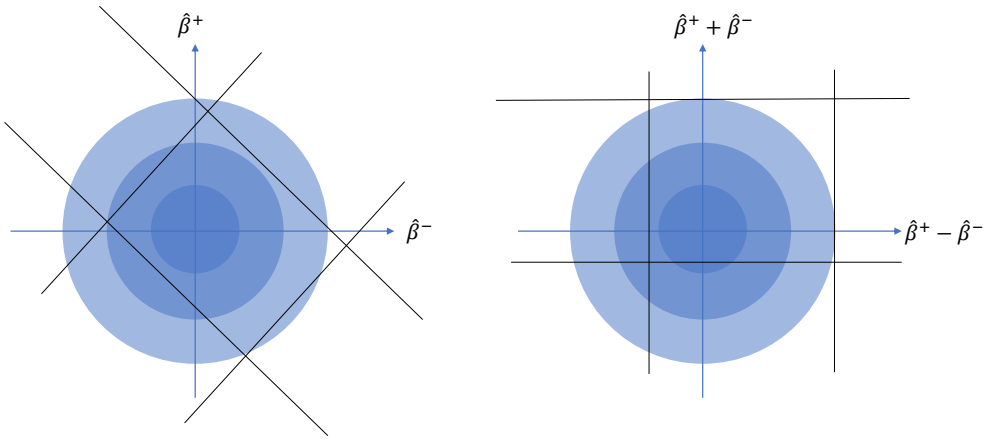

Figure 2: Left: the joint distribution of $\hat{\beta}_j^+$ and $\hat{\beta}_j^-$. Right: Rotate the coordinate system by 45° and obtain the joint distribution $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$.

where

$$\begin{aligned}
\mathcal{V}_0^L(\mathbf{r}) &:= \max_{j:a_{0j}<0} \frac{b_{0j} - (A_0\mathbf{r})_j}{\alpha a_{0j}},\ \mathcal{V}_0^U(\mathbf{r}) := \min_{j:a_{0j}>0} \frac{b_{0j} - (A_0\mathbf{r})_j}{\alpha a_{0j}},\ \mathcal{V}_0^N(\mathbf{r}) := \min_{j:a_{0j}=0} \boldsymbol{b}_{0j} - (A_0\mathbf{r})_j, \\
\mathcal{V}_1^L(\mathbf{r}) &:= \max_{j:a_{1j}<0} \frac{b_{1j} - (A_1\mathbf{r})_j}{\alpha a_{1j}},\ \mathcal{V}_1^U(\mathbf{r}) := \min_{j:a_{1j}>0} \frac{b_{1j} - (A_1\mathbf{r})_j}{\alpha a_{1j}},\ \mathcal{V}_1^N(\mathbf{r}) := \min_{j:a_{1j}=0} b_{1j} - (A_1\mathbf{r})_j,
\end{aligned} \tag{18}$$

with $\alpha = \boldsymbol{\eta}_1^\top \boldsymbol{\eta}_1 = \boldsymbol{\eta}_2^\top \boldsymbol{\eta}_2$, a_{0j}, a_{1j} *are the* j*-th element of* $\mathbf{a}_0, \mathbf{a}_1$ *in (17),* b_{0j}, b_{1j} *are the* j*th element of* $\boldsymbol{b}_0(\mathcal{S}, \mathbf{s})$ *and* $\boldsymbol{b}_1(\mathcal{S}, \mathbf{s})$ *in (13), respectively.*

By the normality of $\mathbf{y}$ and (16), we have $\mathrm{Cov}(\hat{\beta}_j^+ + \hat{\beta}_j^-, \hat{\beta}_j^+ - \hat{\beta}_j^-) = \sigma^2(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)^\top(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) = 0$. $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$ are uncorrelated and hence independent. Since $\mathbf{r}$ is independent of $\boldsymbol{\eta}^\top \mathbf{y}$, it behaves as fixed quantities for the distribution of $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$ conditioning on $\mathcal{S}$. Thus $\hat{\beta}_j^+ + \hat{\beta}_j^- | \{A\mathbf{y} \le \boldsymbol{b}\}$ and $\hat{\beta}_j^+ - \hat{\beta}_j^- | \{A\mathbf{y} \le \boldsymbol{b}\}$ behave like two independent truncated normal random variables. We next use the probability integral transformation to construct a uniform distribution related to $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$.

Theorem 2. *Let* $F_{\mu,\sigma^2}^{[a,b]}$ *denote the CDF of a* $N(\mu, \sigma^2)$ *random variable truncated to the interval* $[a, b]$*, that is,*

$$F_{\mu,\sigma}^{[a,b]}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \tag{19}$$

where Φ *is the CDF of a* $N(0,1)$ *random variable. Then for* $\mathcal{S} \supseteq \mathcal{S}_1$ *and* $j \in \mathcal{S}_0$*, we have*

$$F_{0,\sigma^2}^{[\mathcal{V}_1^L(\mathbf{r}), \mathcal{V}_1^U(\mathbf{r})]}(\hat{\beta}_j^+ + \hat{\beta}_j^-) \mid \hat{\mathcal{S}} = \mathcal{S}, \hat{\mathbf{s}} = \mathbf{s} \sim Unif(0,1) \tag{20}$$

$$F_{0,\sigma^2}^{[\mathcal{V}_0^L(\mathbf{r}), \mathcal{V}_0^U(\mathbf{r})]}(\hat{\beta}_j^+ - \hat{\beta}_j^-) \mid \hat{\mathcal{S}} = \mathcal{S}, \hat{\mathbf{s}} = \mathbf{s} \sim Unif(0,1) \tag{21}$$

and $\mathcal{V}_0^L(\mathbf{r})$, $\mathcal{V}_0^U(\mathbf{r})$, $\mathcal{V}_1^L(\mathbf{r})$ *and* $\mathcal{V}_1^U(\mathbf{r})$ *are defined in (18).*

Denote $U_j = F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ + \hat{\beta}_j^-)$, and $V_j = F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ - \hat{\beta}_j^-)$, where F_{0,σ^2} is the CDF of a $N(0,\sigma^2)$ random variable. Theorem 2 shows that, conditioning on the selection set $(\mathcal{S}, \mathbf{s})$, for $j \in \mathcal{S}_0$, both U_j and V_j follow $N(0,\sigma^2)$ and independent with each other. Therefore, $|U_j| - |V_j|$ have the same distribution with $|V_j| - |U_j|$. We define the mirror statistics M_j as

$$M_j = |F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ + \hat{\beta}_j^-)| - |F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ - \hat{\beta}_j^-)|. \tag{22}$$

For $j \in \mathcal{S}_0$, M_j is symmetrically distributed. When the signal is strong, i.e., $j \in \mathcal{S}_1$, as shown in Lee et al. (2016), the truncation points are far from zero. In such a case, the truncated CDF are close to F_{0,σ^2}, indicating the M_j defined in (22) is close to $|\hat{\beta}_j^+ + \hat{\beta}_j^-| - |\hat{\beta}_j^+ - \hat{\beta}_j^-|$. In practice, the mirror statistics tends to be conservative since we use CDF with mean zero for both null and non-null variables. For non-null variables, the mean for $\hat{\beta}_j^+ + \hat{\beta}_j^-$ and $\hat{\beta}_j^+ - \hat{\beta}_j^-$ before the truncation are $\boldsymbol{\eta}_1^\top \mu$ and $\boldsymbol{\eta}_2^\top \mu$, respectively, where $\mu = \mathbb{E}\mathbf{y}$. We find that replacing μ by a reasonable estimate, such as $\boldsymbol{X}_{\mathcal{S}}\tilde{\beta}$, helps increase the power with little lose of FDR control.

The above properties of M_j hold on the event $\{\hat{\mathcal{S}} = \mathcal{S}, \hat{s} = s\}$. We follow the selected model framework in Tibshirani (2013); Fithian et al. (2014) so that any set $\{\mathcal{S} \supseteq \mathcal{S}_1\}$, i.e. the Lasso has screened successfully. The union of these events is defined as

$$\mathcal{E} = \bigcup_{\mathcal{S} \supseteq \mathcal{S}_1} \bigcup_{s \in \{-1,1\}^{|\mathcal{S}|}} \{\hat{\mathcal{S}} = \mathcal{S}, \hat{s} = s\} = \{\mathcal{S}_1 \subset \hat{\mathcal{S}}\} \tag{23}$$

This event includes all possible selections that selecting all of the nonzero variables while allow false discovered variables. The GM approach is designed to pick out those false discovered ones. Such selection consistency can be guaranteed by the L^q-consistency of Lasso estimate under certain conditions; see (Candès and Plan, 2009; Knight and Fu, 2000; Zhao and Yu, 2006; Van de Geer, 2008; Zhang and Huang, 2008; Meinshausen and Bühlmann, 2006; Meinshausen and Yu, 2009).

In Assumption 1, we suppose the compatibility condition and signal strength assumption in Bühlmann and Van De Geer (2011) hold to guarantee the L^q-consistency in Lemma 4.

Assumption 1. *(a) (Compatibility Condition): Let $\phi_0 > 0$ be a constant. For a $p \times 1$ vector α satisfying $||\alpha_{\mathcal{S}_0}||_1 \leq c_0 ||\alpha_{\mathcal{S}_1}||_1$, we assume*

$$||\alpha_{\mathcal{S}_1}||_1^2 \leq \frac{s_1}{\phi_0^2} \alpha^T \boldsymbol{X}^T \boldsymbol{X} \alpha, \tag{24}$$

where the entry $\alpha_i \in \alpha_{\mathcal{S}_0}$ if $i \in \mathcal{S}_0$, and $\alpha_{\mathcal{S}_1} = \alpha \setminus \alpha_{\mathcal{S}_0}$, $s_1 := |\mathcal{S}_1|$, c_0 is a constant depending on the choice of λ_n, and ϕ_0 is the compatibility constant.

(b) (Signal Strength): $\min_{j \in \mathcal{S}_1} |\beta_j| > \frac{64\sigma s_1}{\phi_0^2}\sqrt{\frac{\log(p)}{n}}$.

Lemma 4. *Suppose that Assumption 1 holds. Consider the Lasso with regularization parameter* $\lambda = 4\sigma\sqrt{\log(p)n}$. *Then we have*

$$P(\mathcal{S}_1 \subset \hat{\mathcal{S}}) \geq 1 - \frac{2}{p},$$

where $\hat{\mathcal{S}}$ *is the index set of the nonzero entries in the Lasso estimator* $\tilde{\boldsymbol{\beta}}$ *in (11).*

Lemma 4 follows directly from the $||\cdot||_1$ convergence result in Bühlmann and Van De Geer (2011), Theorem 6.1 and the signal strength condition in Assumption 1(b). Based on this lemma, the event $\mathcal{E} = \{\mathcal{S}_1 \subset \hat{\mathcal{S}}\}$ holds with high probability, that is, the Lasso estimator can select all of the non-zero variables while allow false discovered variables. The GM approach is designed to pick out those false discovered ones; and a symmetric statistics M_j constructed on the set $\hat{\mathcal{S}}$ is sufficient to achieve this goal.

Algorithm 2 Gaussian mirror algorithm for Lasso.

1. Fit Lasso with respect to the original design matrix $\boldsymbol{X}$, i.e.,

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda_n||\boldsymbol{\beta}||_1$$

   by cross validation.

2. Parallel FOR $j \in 1, \ldots, \hat{\mathcal{S}}$:

   (a) Generate $\boldsymbol{z}_j$ from Gaussian distribution with mean zero and covariance matrix $\boldsymbol{I}_n$.

   (b) Calculate debiased c_j via (14).

   (c) Calculate the mirror statistics

$$M_j = |F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ + \hat{\beta}_j^-)| - |F_{0,\sigma^2}^{-1} F_{0,\sigma^2}^{[\mathcal{V}^-(\mathbf{r}),\mathcal{V}^+(\mathbf{r})]}(\hat{\beta}_j^+ - \hat{\beta}_j^-)| \tag{25}$$

   END parallel FOR loop

3. Calculate the cutoff to control FDR at target q.

$$\tau_q = \min_t \Big\{ \frac{\#\{j \mid M_j \leq -t\}}{\#\{j \mid M_j \geq t\} \vee 1} \leq q \Big\}$$

4. Output the index of the selected variables: $\hat{\mathcal{S}}_1 = \{j \,|\, M_j \geq \tau_q\}$.

Theorem 3. *Let* M_j *be the mirror statistics defined in (22), which can be computed using Algorithm*

2 in this section, with the GM design in Definition 2. We have

$$P(M_j \le -t \mid \boldsymbol{z}_j) = P(M_j \ge t \mid \boldsymbol{z}_j),\ \forall t > 0.$$

for $j \in \mathcal{S}_0$ with probability $1 - 2/p$.

Based on Theorem 3, we show that the symmetric property of M_j holds for finite sample size for any $j \in \mathcal{S}_0$. We use $\#\{j \,|\, M_j \le -t\}$ as an (over)-estimate of the number of false positive set, and define an estimate of FDP as

$$\widehat{\mathrm{FDP}}(t) \triangleq \frac{\#\{j \,|\, M_j \le -t, j \in \hat{\mathcal{S}}\}}{\#\{j \,|\, M_j \ge t, j \in \hat{\mathcal{S}}\} \vee 1}. \tag{26}$$

Algorithm 2 illustrates the detailed procedure from the Gaussian mirror design to false discovery control. Similar to the GM algorithm for the OLS in Section 2.2, we compute M_j in parallel for $j = 1, \ldots, p$.

4 Theory for FDR Control with Gaussian Mirrors

Section 2 introduces a Gaussian mirror-based method to estimate the FDP. Controlling FDR is useful when the number of features is large. For example, modern gene expression studies usually involve thousands of genes; and genome-wide association studies (GWAS) routinely examine tens of thousands to millions of single nucleotide polymorphisms (SNPs). It is of practical importance to select a set of small number of significant genes or SNPs for follow-up experimental validations and testings, which requires the scientist to have a reliable estimation and control of the FDR so as to control the experimental cost. In this section, we show that the FDR can indeed be controlled at any given level q asymptotically by using Gaussian mirrors.

Without loss of generality, we use M_j to indicate both the one defined in (4) and the one defined in (22). The key is to design the Gaussian mirror appropriately such that for $j \in \mathcal{S}_0$, $P(M_j \le -t \mid \boldsymbol{z}_j) = P(M_j \ge t \mid \boldsymbol{z}_j)$, $\forall t > 0$. When we select the j-th variable if $M_j \ge t$, then the FDP satisfies

$$\frac{\#\{j \in \mathcal{S}_0 : M_j \ge t\}}{\#\{j : M_j \ge t\} \vee 1} \approx \frac{\#\{j \in \mathcal{S}_0 : M_j \le -t\}}{\#\{j : M_j \ge t\} \vee 1} \le \frac{\#\{j : M_j \le -t\}}{\#\{j : M_j \ge t\} \vee 1}, \tag{27}$$

where the last term is the $\widehat{FDP}(t)$, an (over-)estimate of the FDP defined in (5). As shown in Algorithms 1 and 2, a data-driven threshold $\tau_q > 0$ is chosen in (6) as the smallest value such that $\widehat{FDP}(\tau_q) \le q$. To proceed, we need the following weak dependence assumption on the mirror statistics M_j's for $j = 1, \cdots, p$, as $p \to \infty$.

Assumption 2. *(a) Let $\mathcal{T}$ be a subset of $\{1, 2, \cdots, p\}$, and $\alpha < 3/2$ be a constant, the mirror statistics M_j's satisfy*

$$Cov\Big(\sum_{j\in\mathcal{T}} 1(M_j \geq t), \sum_{k\in\mathcal{T}} 1(M_k \geq t)\Big) \leq C|\mathcal{T}|^{\alpha}, \ \forall\, \mathcal{T}, t$$

where $1(\cdot)$ is an indicator function, C is an absolute constant;

(b) Let $p_0 = |\mathcal{S}_0|$, and $p_1 = |\mathcal{S}_1|$ for OLS; $p_0 = |\mathcal{S}_0 \cap \hat{\mathcal{S}}|$, $p_1 = |\hat{\mathcal{S}}|$ for Lasso post-selection, we assume

$$\lim_p \frac{p_0}{p_0 + p_1} = \pi_0 > 0, \quad \lim_p \frac{p_1}{p_0 + p_1} = \pi_1 = 1 - \pi_0 > 0.$$

Assumption 2(a) quantifies the pairwise correlation among all the M_j's. When the M_j's are pairwise independent, the left hand side is $O(|\mathcal{T}|)$. When the M_j's are perfectly correlated, the left hand side is $O(|\mathcal{T}|^2)$. Here we assume $\alpha < 3/2$ to incorporate cases where the M_j's are moderately correlated. Empirically, we observe that pairwise correlations among the M_j's are always weaker than those among the $\hat{\beta}_j$'s due to the random perturbation introduced by $c_j \boldsymbol{z}_j$, which provides an explanation why the GM tends to be more powerful than both the BH and knockoff when the covariates are moderately to highly correlated. Assumption 2(b) essentially requests that the number of null variables casts a constant proportion of the total number of variables.

Lemma 5. *Define a few quantities: $V(t) = \#\{j : j \in \mathcal{S}_0, M_j \leq -t\}, V'(t) = \#\{j : j \in \mathcal{S}_0, M_j \geq t\}, U(t) = \#\{j \in \mathcal{S}_1, M_j \leq -t\}, W(t) = \#\{j \in \mathcal{S}_1, M_j \geq t\}$ and $R(t) = \{j : M_j \geq t\}$. Define $G_0(t) = \lim_p \frac{1}{p_0}\sum_{j\in\mathcal{S}_0} \mathbb{E}1(M_j \leq -t)$, $G_1(t) = \lim_p \frac{1}{p_1}\sum_{j\in\mathcal{S}_1} \mathbb{E}1(M_j \geq t)$, and $G_2(t) = \lim_p \frac{1}{p_1}\sum_{j\in\mathcal{S}_1} \mathbb{E}1(M_j \leq -t)$. Suppose Assumption 2 holds and both $G_0(t)$, $G_1(t)$, and $G_2(t)$ are continuous function, then*

$$\limsup_t \left|\frac{V(t)}{p_0} - G_0(t)\right| = \limsup_t \left|\frac{V'(t)}{p_0} - G_0(t)\right| = 0,$$

$$\limsup_t \left|\frac{U(t)}{p_1} - G_2(t)\right| = \limsup_t \left|\frac{W(t)}{p_1} - G_1(t)\right| = \limsup_t \left|\frac{R(t)}{p} - \pi_0 G_0(t) + \pi_1 G_1(t)\right| = 0,$$

almost surely.

The proof of Lemma 5 is provided in the Appendix 8. With the aid of this lemma, we have the following result.

Theorem 4. *Let M_j be the test statistics calculated using the OLS. For any given q-level, we choose $\tau_q > 0$ according to (6). If Assumption 2 holds, then, as $p \to \infty$,*

$$\mathbb{E}\left[\frac{\#\{j : j \in \mathcal{S}_0, \text{ and } j \in \hat{\mathcal{S}}_1\}}{\#\{j : j \in \hat{\mathcal{S}}_1\} \vee 1}\right] \leq q. \tag{28}$$

Proof. First, we show $\tau_q \le C$ for some sufficiently large constant C in probability 1. For any $\epsilon > 0$, let t^* be chosen such that

$$\frac{\pi_0 G_0(t^*) + \pi_1 G_2(t^*)}{\pi_0 G_0(t^*) + \pi_1 G_1(t^*)} < q - \epsilon.$$

When p is sufficiently large, based on Lemma 5, we have

$$\left| \frac{V(t^*) + U(t^*)}{R(t^*) \vee 1} - \frac{\pi_0 G_0(t^*) + \pi_1 G_2(t^*)}{\pi_0 G_0(t^*) + \pi_1 G_1(t^*)} \right| < \epsilon/2,$$

Then

$$\left| \frac{V(t^*) + U(t^*)}{R(t^*) \vee 1} \right| \le q - \epsilon/2.$$

Note that τ_q is chosen such that

$$\tau_q = \arg\min_t \left\{ t : \frac{V(t) + U(t)}{R(t) \vee 1} \le q \right\},$$

implying that $\tau_q < t^*$.

Note that for a sufficient large constant $C > t^*$, by Lemma 5,

$$\lim \sup_{0 \le t \le C} \left| \frac{V(t)}{R(t) \vee 1} - \frac{\pi_0 G_0(t)}{\pi_0 G_0(t) + \pi_1 G_1(t)} \right| = 0.$$

Then we have

$$\begin{aligned} & \lim \sup_{0 \le \tau_q \le C} \mathbb{E} \frac{\sum_{i \in \mathcal{S}_0} 1(M_i \ge \tau_q)}{\sum_i 1(M_i \ge \tau_q) \vee 1} \le \mathbb{E} \lim \sup_{0 \le \tau_q \le C} \frac{\sum_{i \in \mathcal{S}_0} 1(M_i \ge \tau_q)}{\sum_i 1(M_i \ge \tau_q) \vee 1} \\ = \; & \mathbb{E} \lim \sup_{0 \le \tau_q \le C} \frac{V'(\tau_q)}{R(\tau_q) \vee 1} = \mathbb{E} \lim \sup_{0 \le \tau_q \le C} \frac{V(\tau_q)}{R(\tau_q) \vee 1} \le q. \end{aligned}$$

The first inequality is based on the Fatou's lemma and the last inequality is based on the definition of τ_q. □

The next theorem provides a FDR control for Gaussian mirror algorithm for Lasso.

Theorem 5. *Let M_j be the test statistics calculated using the Algorithm 2. For any given q-level, we choose $\tau_q > 0$ according to (6). If Assumptions 1 and 2 hold, then, as $p \to \infty$,*

$$\mathbb{E}\left[\frac{\#\{j : j \in \mathcal{S}_0, \text{ and } j \in \hat{\mathcal{S}}_1\}}{\#\{j : j \in \hat{\mathcal{S}}_1\} \vee 1} \right] \le q + o(1). \tag{29}$$

When conditioning on the event that $S_1 \subset \hat{\mathcal{S}}$, we can follow the proof of Theorem 4. Combining this with Lemma 4, we have this theorem.

5 Estimating the Number of False Discoveries

The GM approach introduced in Section 2 provides a way to select significant features with controlled FDR. Here we address a problem closely related to FDR controls but may be of more direct interest to some practitioners. Suppose the scientists are only allowed to explore no more than 100 selected features due to limited budget, then the following questions are of immediate concerns: (1) How should they select the list of top 100? (2) How many false discoveries (FDs) do they expect to have? (3) Can statisticians provide an uncertainty measure for the estimated FDs?

The first question is easy to address based on the mirror statistics $\{M_j\}$ $(j = 1, \cdots, p)$. We order the mirror statistics decreasingly as $M_{(1)} \geq \cdots \geq M_{(p)}$ and choose the top k features, as those corresponding to the set of the top-k mirror statistics. Denote the selected set as $\mathcal{I}_k$, i.e.,

$$\mathcal{I}_k = \{j : M_j \in \{M_{(1)}, \ldots, M_{(k)}\}\},$$

where $k \leq \#\{j \mid M_j > 0\}$ since the GM procedure does not select variables with negative mirror statistics. Let $FD(k) = |\mathcal{I}_k \cap \mathcal{S}_0|$ denote the true number of FDs in the top-k list, which is a random variable with its randomness arising from $\boldsymbol{X}$, $\{\epsilon_i\}$, and $\{\boldsymbol{z}_j\}$. The expected number of FDs, $\mathbb{E}[FD(k)]$, is a more stable target for estimation. Since the mirror statistics is distributed symmetrically for the null variables, we estimate $\mathbb{E}[FD(k)]$ as

$$\widehat{FD}(k) = \#\{M_j < -M_{(k)}\}. \tag{30}$$

Theorem 6 below shows that the error bound between $\widehat{FD}(k)$ and $\mathbb{E}[FD(k)]$ is $o_p(k)$. As k increases, the error bound also increases although the relative error gets smaller. Theorem 7 states that $\widehat{FD}(k)$ is an unbiased estimator of $\mathbb{E}[FD(k)]$ with high probability.

Theorem 6. *Suppose that Assumptions 1 and 2 hold, $p/n \to \infty$, $k \leq \#\{j \mid M_j > 0\}$, and $k/p_1 = O(1)$. Let M_j be the mirror statistics calculated from Algorithm 2. We have*

$$\lim_{n\to\infty} P\left(\frac{1}{k}\left|\widehat{FD}(k) - \mathbb{E}[FD(k)]\right| > \epsilon\right) = 0.$$

for any $\epsilon > 0$.

Theorem 7. *Suppose that Assumptions 1 holds, $k \leq \#\{j \mid M_j > 0\}$. Let M_j be the mirror statistics calculated from Algorithm 2. We have*

$$P\left(\mathbb{E}[\widehat{FD}(k)] = \mathbb{E}[FD(k)]\right) > 1 - \frac{2}{p}.$$

Next, we describe a nonparametric bootstrap method to construct a confidence interval for $\mathbb{E}[FD(k)]$. The method starts by fitting the regression model based on the original design matrix

using either the least squares (if $p < n$) or the Lasso method, and obtaining the fitted values $\hat{\boldsymbol{y}} = (\hat{y}_1, \cdots, \hat{y}_n)$ as well as the residuals $\boldsymbol{\gamma} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. Then, for $b = 1, \ldots, B$, we generate independently the b-th "bootstrap sample" $\boldsymbol{y}^{(b)} = \hat{\boldsymbol{y}} + \boldsymbol{\gamma}^{(b)}$, where the $\boldsymbol{\gamma}^{(b)}$ is drawn randomly from $\boldsymbol{\gamma}$ with replacement. With the bootstrapped "observations" $\boldsymbol{y}^{(b)}$, we calculate the mirror statistics $\{M_1^{(b)}, \ldots, M_p^{(b)}\}$ using Algorithm 1 (Algorithm 2 for Lasso). The B sets of bootstrap mirror statistics are denoted as $\{M_1^{(b)}, M_2^{(b)}, \cdots, M_p^{(b)}\}_{b=1}^B$, which give rise to a set of B bootstrap estimates of $FD(k)$: $\mathcal{B}_{FD} = \{\widehat{FD}^{(1)}(k), \ldots, \widehat{FD}^{(B)}(k)\}$. A bootstrap confidence interval for $\mathbb{E}[FD(k)]$ can be constructed as the upper and lower $\alpha/2$ quantiles of the sample $\mathcal{B}_{FD}$, denoted as $\widehat{FD}_{(\alpha/2)}(k)$ and $\widehat{FD}_{(1-\alpha/2)}(k)$, respectively. We may also first re-center the set $\mathcal{B}_{FD}$ to have mean $\widehat{FD}(k)$ and then use the corresponding quantiles. If a budget-sensitive domain scientist is only interested in a $(1-\alpha)100\%$ upper confidence bound of $\mathbb{E}[FD(k)]$, then she/he can use $\widehat{FD}_{(1-\alpha)}(k)$.

Algorithm 3 Bootstrap distribution of M_j and $\widehat{FD}(k)$.

1. Parallel FOR $b = 1, \cdots, B$:

   (a) For low-dimensional cases, fit linear regression model via OLS, i.e., minimizing (3); For high dimensional cases, fit Lasso via minimizing the penalized least squares in (11). Let $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$ be the fitted values, and $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (r_1, r_2, \cdots, r_n)$ be the residuals.

   (b) Sample from $\boldsymbol{r}_j$ with replacement to get $\boldsymbol{r}^{(b)} = (r_1^{(b)}, r_2^{(b)}, \cdots, r_n^{(b)})$. Let $\boldsymbol{y}^{(b)} = \hat{\boldsymbol{y}} + \boldsymbol{r}^{(b)}$.

   (c) Apply Algorithm 1 (low-dimensional case) or Algorithm 2 (high-dimensional case) by replacing $\boldsymbol{y}$ with $\boldsymbol{y}_j^{(b)}$, and calculate the mirror statistics M_j^b for $j = 1, \ldots, p$ accordingly.

   End Parallel FOR.

2. Output the bootstrap estimate $\{M_1^b, M_2^b, \cdots, M_p^b\}_{b=1}^B$.

3. For $b = 1, \ldots, B$, we calculate the estimate of the number of false discoveries based on $\{M_1^b, M_2^b, \cdots, M_p^b\}$ as
$$\widehat{FD}^b(k) = \#\{M_j^b < -M_{(k)}^b\}.$$
where $\{M_{(1)}^b, \ldots M_{(p)}^b\}$ are the decreasingly ordered mirror statistics.

4. Construct the confidence interval of $\mathbb{E}[FD(k)]$ as $[\widehat{FD}_{(\alpha/2)}(k), \widehat{FD}_{(1-\alpha/2)}(k)]$ and upper confidence interval as $[0, \widehat{FD}_{(1-\alpha)}(k)]$.

6 Numerical Studies

6.1 Low dimensional scenarios ($p < n$)

The first method is BH Benjamini and Hochberg (1995), which controls FDR through finding a data adaptive threshold for the p-values of the regression coefficients. The second method is the knockoff Barber and Candès (2015), which introduces the knockoff filter to control the FDR in the sparse linear model whenever there are at least as many observations as variables. The third method is the model-X knockoff Candes et al. (2018), which extends the knockoff procedure to high-dimensional settings. For the model-X knockoff, We construct the knockoffs with both known covariance matrix (model-X) and a second-order estimate of the covariance matrix (model-X-est). Also, we observe that the knockoffs constructed based on the known covariance matrix becomes exteme conservative for the constant correlation setting with correlation coefficient larger than 0.5. We implement a modified verison of model-X knockoff construction (model-X-fix) which significantly increases the power. We consider two scenarios in simulations: $p < n$ and $p \geq n$. In all simulation settings, we set the targeted value of the FDR as $q = 0.1$.

We simulate linear regression models with $n = 1000$ and $p = 300$. For the GM approach, we calculate the mirror statistics based on Algorithm 1. For the BH method, we calculate the z-statistics $z_1, \ldots, z_p$ for the OLS estimates, i.e., $z_j = \hat{\beta}_j/(\sigma\sqrt{(X^TX)^{-1}_{jj}})$. The j-th variable is selected if $|z_j| > \tau_q^{BH}$, where τ_q^{BH} is chosen as

$$\tau_q^{BH} = \min_t \left\{ p\frac{P(|N(0,1)| \geq t)}{\#\{j\,||z_j| \geq t\}} \leq q \right\}. \tag{31}$$

The knockoff and model-X knockoff statistics are calculated based on the Lasso estimators. For different methods, we evaluate the FDR and the selection power based on 100 replications. The selection power is calculated as the ratio between the number of the correctly selected covariates and the true nonzero covariates. In each of the following settings, we randomly set 240 coefficients of $\boldsymbol{\beta}$ as zero and generate the remaining 60 nonzero coefficients independently from $N(0, 20/\sqrt{n})$. The response variable $\boldsymbol{y}$ is generated according to e.q (1) with $\sigma = 1$. The design matrix are composed of i.i.d. rows with each row generated from $N(0, \Sigma)$.

(i) Power decay correlation The covariance matrix $\boldsymbol{\Sigma}$ is autoregressive, i.e., each of its element $\sigma_{ij} = \rho^{|i-j|}$, and we take $\rho = 0, 0.2, 0.4, 0.6, 0.8$, respectively. Figure 3 (a1-a2) show the box-plot of the FDPs and the selection power for the five methods. As shown in Figure 3 (a1), the FDPs of the four methods are all around the targeted value 0.1. Figure 3 (a2) shows that the GM method has the highest power among the four methods; and the gap between GM and the other methods gets larger as ρ increases. Particularly, when ρ increases to 0.8, the power of the knockoff method decreases dramatically with some extreme cases of having no rejections at all.

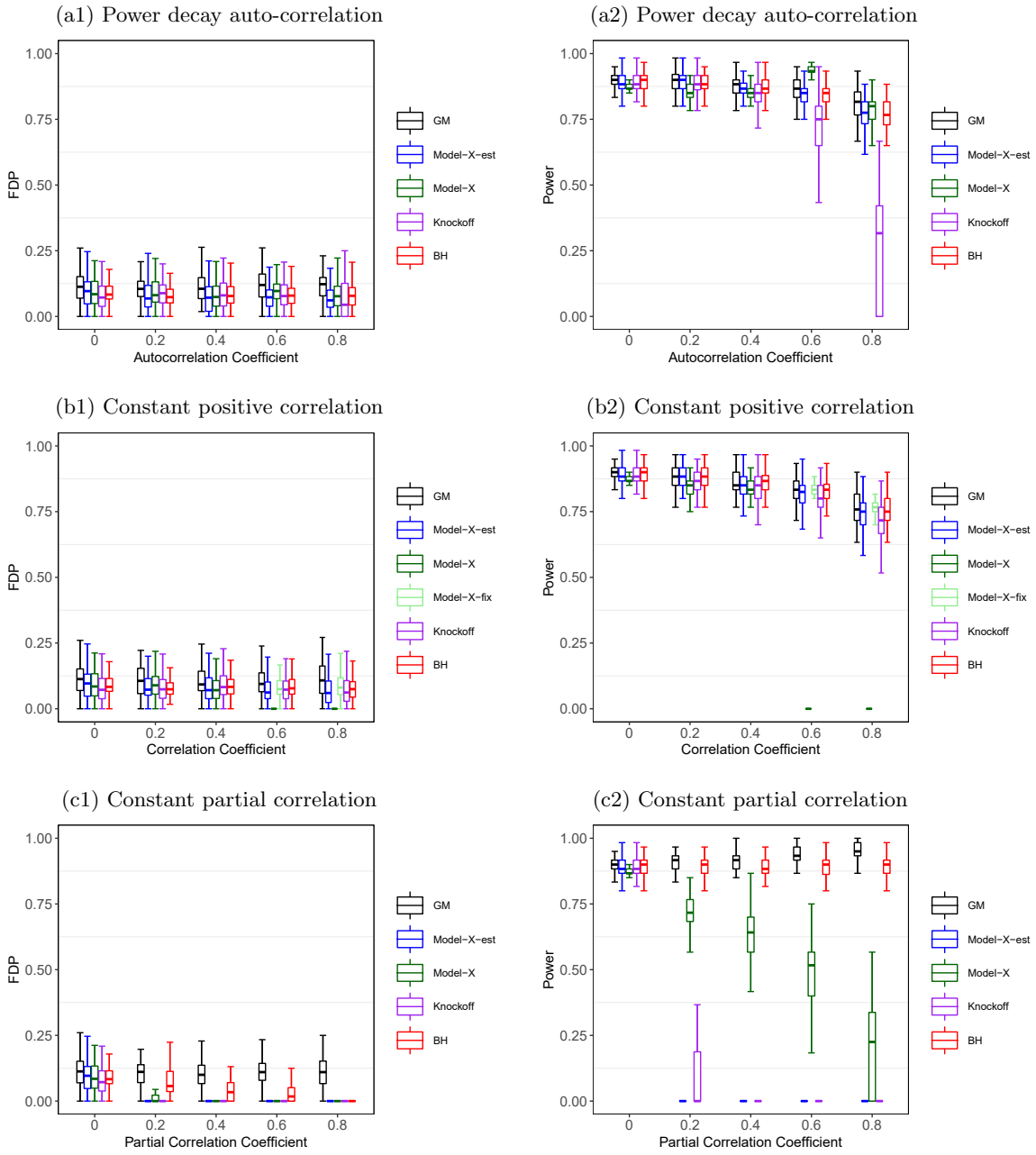

Figure 3: The FDR and power performances for low-dimensional cases. Notations Model-X-est, Model-X, Model-X-fix, knockoff refer to the Model-X knockoff with estimated covaraince matrix, Model-X knockoff with known covariance matrix, modified Model-X knockoff with known covariance matrix, and knockoff, respectively. The upper and lower hinges of the box correspond to the first and third quartiles. The interval is calculated via the mean of the FDP $\pm$ the sample standard deviation of the FDP.

(ii) Constant positive correlation. Here we let $\sigma_{ij} = \rho^{1(i\neq j)}$, with $1(\cdot)$ being an indicator function.The simulation results are summarized in Figure 3 (b1-b2). In this case, the partial correlation of $\boldsymbol{X}$ is low, implying that the correlation among the $\boldsymbol{\beta}$ is small and the test statistics are weakly correlated. The BH method produces the most stable FDPs and maintained a high power. Model-X-est and the GM method come in a close second, obtaining results almost indistinguishable from that of the BH.

When $\rho \geq 0.5$, the power of the Model-X drops to near-zero, which is due to a numerical issue in the original Model-X software. The Model-X procedure generates knockoffs from $N(\mu, V)$, where V is obtained via solving a semi-definite programming problem. When $\rho >= 0.5$, V is nearly a rank-one matrix, which results in high collinearity among the generated knockoffs and significantly reduces the power. Dongming Huang and Lucas Jensen (personal communication) suggested a simple remedy, which is to project each variable onto the orthogonal space of the first principal component of Σ and add a small perturbation as

$$\mathbf{x}_j^{new} = \mathbf{x}_j - \frac{1}{\alpha}\boldsymbol{X}\Sigma^{-1}\mathbf{1}_p + \frac{1}{\sqrt{\alpha}}\boldsymbol{z}_j$$

where $\alpha = (\rho - \rho^2\mathbf{1}_p^T\Sigma^{-1}\mathbf{1}_p)^{-1}$ and $\boldsymbol{z}_j \overset{iid}{\sim} N(0, I_n)$. Then, $\boldsymbol{X}^{new} = (\boldsymbol{x}_1^{new}, \ldots, \boldsymbol{x}_p^{new})$ serves a new design matrix for constructing knockoffs via the standard Model-X procedure with the covariance matrix I_p. This modification significantly increases the power of Model-X knockoff (as shown in Figures 3(b1-b2) and 4(b1-b2)), but it only works for this case, where each $\mathbf{x}_j$ can be represented as the sum of a common latent Gaussian factor and an independent term. Although power decreases for all the methods when ρ increases, the GM method appears to be least affected.

(iii) Constant partial correlation. In this setting, the precision matrix $Q = \boldsymbol{\Sigma}^{-1}$ has constant off-diagonal elements, i.e., $q_{ij} = \tau^{1(i\neq j)}$. The correlation among the $\boldsymbol{X}$ can be very small when τ is large. For example, when $\tau = 0.6$, the off-diagonal entries of $\boldsymbol{\Sigma}$ is about -0.0083 and the diagonal entry is 2.4917. The comparison results are reported in Figure 3(c1-c2). Both the GM and BH methods work well in terms of the FDP and power, while GM is slightly more powerful. When $\tau > 0$, both knockoff and Model-X-est (using estimated covariance matrix) are extremely conservative. For instance, when $\tau = 0.6$, both the FDP and power of these two methods drop to zero even though the pairwise correlations among the $\boldsymbol{x}_j$'s are small. The Mode-X with known covariance matrix works better, the power still decreaes fast as ρ increases, much worse than both the GM and the BH methods.

6.2 High dimensional scenarios ($p \geq n$)

We consider the high dimensional case with $p = 1000$ and $n = 300$, in which Algorithm 2 is used for the GM method. Similar to Candes et al. (2018), we implement the marginal BH (henceforth,

BH-ma), which is based on the p-value of the marginal regression between $\boldsymbol{y}$ and $\boldsymbol{x}_j$, $j = 1, \cdots, p$. We also consider the data splitting BH (henceforth, BH-ds), which employs Lasso to select variables based on the first half of the data, and applies the BH method based on the p-values of the OLS fitting of the selected variables using the second half of the data. For the Model-X knockoff, we use both the default setting Gaussian knockoffs with known covariance matrix and the second-order knockoffs with estimated covariance matrix as implemented in the R package *knockoff*. Note that the original knockoff method can not be applied to the high dimensional case when $p \geq n$.

We consider the same three correlation structures of the design matrix $\boldsymbol{X}$ as in Section 6.1. In each setting, we set 940 coefficients of $\boldsymbol{\beta}$ as zero and generate the remaining 60 coefficients from $N(0, 20/\sqrt{n})$. The response variable $\boldsymbol{y}$ is generated according to e.q (1) with $\sigma = 1$. We evaluate the FDPs and powers of the methods based on 100 replications, with the target FDP set at $q = 0.1$.

(i) Power decay correlation. As shown in Figure 4 (a1), the FDPs of the GM method are well controlled around 0.1 in all cases. The FDPs of Model-X-est, Model-X, BH-ds are well controlled when $\rho \leq 0.6$. When $\rho \geq 0.8$, the FDPs of Model-X-est are inflated to 0.23. This inflation effect is also observed in Candes et al. (2018). The FDPs of BH-ms start to get significantly inflated when $\rho \geq 0.4$. We see in Figure 4 (a2) that the GM and two Model-X methods have very comparable powers, whereas the two BH methods do not perform as well.

(ii) Constant Positive Correlation. As shown in Figure 4(b1), GM, BH-ds, and Model-X/Model-X-fix all control their FDPs properly around 0.1. Model-X-est is overly conservative with the observed FDP near zero when $\rho \geq 0.2$, whereas BH-ma fails to control the FDR especially when $\rho \geq 0.2$ mainly because of strong correlations among the covariates. When ρ is larger than 0.5, we apply the same algorithmic modification as in the low dimensional scenario for Model-X, denoted as Model-X-fix, which improves the power. Figure 4(b2) shows the power comparisons. The power of Model-X-est knockoff decreases rapidly as ρ increases, getting to zero when $\rho \geq 0.6$. In contrast, Model-X/Model-X-fix maintains a better power than Model-X-est. The power of BH increases dramatically with ρ increases, at the expense of an equally rapidly growing FDP. The powers of GM maintain the most stable power, decreasing only moderately from 0.75 as ρ increases.

(iii) Constant partial correlation. As shown in Figure 4(c1), all methods have their FDPs well under control in all cases, except that BH-ma somehow has its FDP greatly inflated when $\rho = 0.8$ (Figure 4 (c1)). Both GM and Model-X-est perform well and comparably in terms of both the FDR control and power, with GM having a slight advantage, while BH-ma, BH-ds and Model-X have significantly lower powers than GM and Model-X-est for $\rho > 0$.

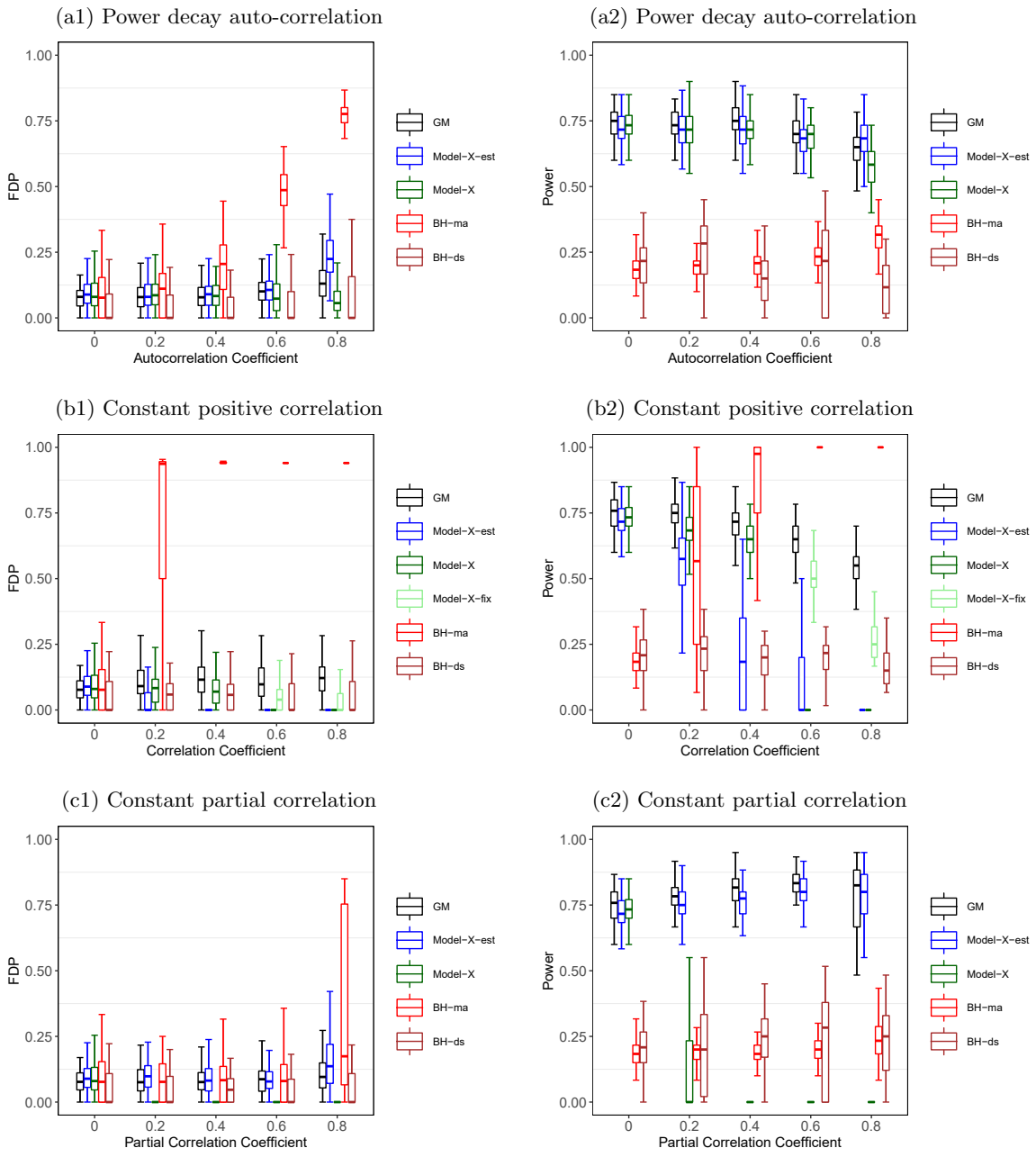

Figure 4: The FDR and power performance for high-dimensional scenarios. Notations are the same as those in Figure 3. Note that Model-X-fix, which fixed a previous numerical error of the knockoff package, was applied only to case (b) for $\rho \geq 0.6$.

6.3 Simulation studies based on a population genetics data set

Due to the recent biotechnology revolution, genome wide association studies (GWAS) have become an attractive tool for genetic research. In such studies, researchers examine a genome-wide set (tens of thousands to millions) of genetic variants of a group of individuals, hopefully randomly selected from the target population, to see if any variant is associated with a phenotype of interest. Genetic variants are usually in the form of single nucleotide polymorphisms (SNPs), and are often used as covariates in a linear or logistic regression model to fit the observed phenotype, but with its main goal being variable selections. Since the SNPs often take on only three values, $\{0, 1, 2\}$ (representing 0, 1, or 2 minor allele mutations, respectively), the design matrix of any SNP-based regression is clearly non-Gaussian.

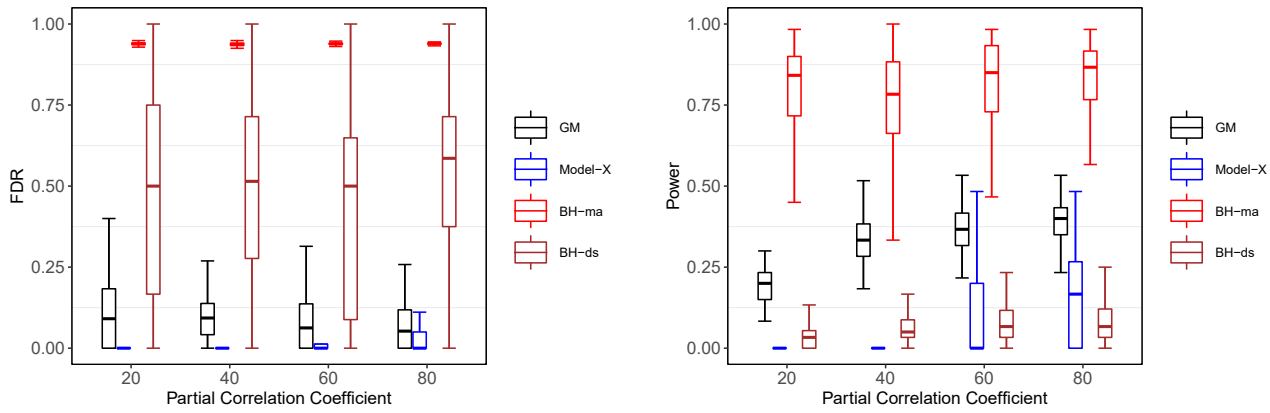

Figure 5: The FDP and power for GM and Model-X knockoff for the GWAS-based design matrix designs. The upper and lower "hinges" of each box correspond to the first and third quartiles of the 100 replications.

We consider a panel of 292 tomato accessions in Bauchet et al. (2017), which is publicly available at ftp://ftp.solgenomics.net/manuscripts/Bauchet_2016/ and includes breeding materials (specimens) characterized by $> 11,000$ SNPs. Here we are interested in examining the FDP and power of GM, BH, and Model-X knockoff by using this real-data set to create realistic design matrices. Specifically, we randomly select 1000 SNPs as $\boldsymbol{X}$ and randomly generate 60 nonzero regression coefficients from $N(0, c^2/n)$, where c ranges from 20 to 80. The response variable $\boldsymbol{y}$ is generated from e.q (1) with standard Gaussian noise. We set the target FDP as 10% and calculate the FDP and the averaged power based on 100 independent replications.

As shown in Figure 5, we observe that GM controls the FDR at the target 10% quite precisely, while Model-X is too conservative with near-zero FDP. As a consequence, Model-X has a significant lower power than GM. The BH methods (both the marginal regression and the data splitting

version), on the other hand, completely lose the FDR control because of high correlations among the randomly selected SNPs. As a consequence, the high power of BH-ma is not scientifically meaningful in this case.

6.4 Empirical results on estimating the expected number of FDs

In this part, we assess the performance of the GM method for estimating the expected number of FDs in a top-k list, denoted as $FD(k)$, as well as the coverage probability of its bootstrap confidence interval proposed in Section 5. We consider both low dimensional ($n = 1000, p = 300$) and high dimensional scenarios ($n = 300, p = 1000$). For both scenarios, we randomly generate 60 nonzero coefficients independently from $N(0, 20/n)$. The design matrix follows the same covariance settings as those described in Sections 6.1 and 6.2 with $\rho = 0.2$. The response y is generated according to e.q (1) with $\sigma = 1$. We repeat 100 times for each setting.

In each replication, we obtain the estimate of $FD(k)$ with $k \in (50, 70)$. For the rth replication ($1 \leq r \leq 100$), we calculate $\widehat{FD}^{(r)}(k)$ following (30), and record the underlying true number of false discoveries as $FD^{(r)}(k)$. We use the sample average $\tilde{\mathbb{E}}[FD(k)] := \frac{1}{100}\sum_{r=1}^{100} FD^{(r)}(k)$ as an approximation to $\mathbb{E}[FD(k)]$ and use $\tilde{\mathbb{E}}[\widehat{FD}(k)] := \frac{1}{100}\sum_{r=1}^{100} \widehat{FD}^{(r)}(k)$ as an approximation of $\mathbb{E}[\widehat{FD}(k)]$, which should ideally track the value of $\mathbb{E}[FD(k)]$.

By using Algorithm 3 with $B = 200$, we calculate the empirical coverage probability of the proposed bootstrap confidence interval for $\mathbb{E}[FD(k)]$ as

$$\frac{1}{100}\#\{r : \tilde{\mathbb{E}}[FD(k)] \in [\widehat{FD}_{(\alpha/2)}(k), \widehat{FD}_{(1-\alpha/2)}(k)]\}$$

where $\widehat{FD}^{(r)}_{(\alpha/2)}(k)$ and $\widehat{FD}^{(r)}_{(1-\alpha/2)}(k)$ are the $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution of $\widehat{FD}(k)$ in the rth replication. In addition, we evaluate the empirical coverage probability of the bootstrap $(1-\alpha)100\%$ upper bound

$$\frac{1}{100}\#\{r : \tilde{\mathbb{E}}[FD(k)] \in [0, \widehat{FD}_{(1-\alpha)}(k)]\}.$$

We set $\alpha = 0.05$ for all simulation cases.

The scatter plots in Figures 6 (a1,b1,c1) show the $\widehat{FD}(k)$ against $FD(k)$, the true number of false discoveries, in each replication of low dimensional cases with k ranging from 50 to 70. The closer the point approaches the diagonal line, the closer the two numbers are. As k increases, the estimate tends to be more dispersed, which is expected since the error bound between $\widehat{FD}(k)$ and $FD(k)$ increases as k increases. In fact, $FD(k)$ is also more variable around its mean $\mathbb{E}(FD(k))$ as k increases. Figure 6 (a2,b2,c2) report the coverage probabilities of the confidence interval and upper confidence interval for $\mathbb{E}[FD(k)]$ in all the settings. The coverage probabilities are above 0.95 when

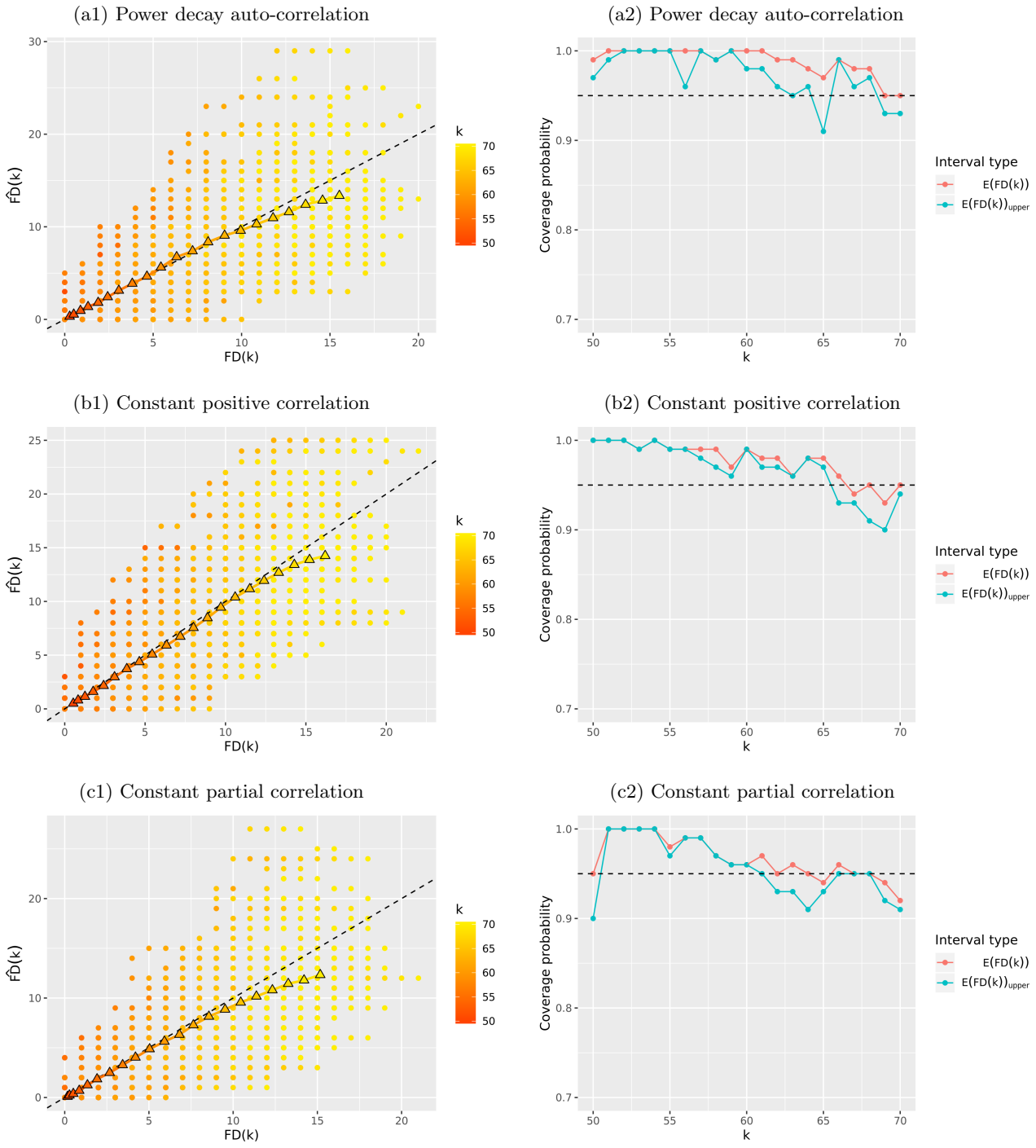

Figure 6: Low dimensional scenario with (n=1000, p=300). Left panels: scatter plots of $\widehat{FD}(k)$ versus $FD(k)$ for $k \in (50, 70)$, under three correlation structures of the design matrix, respectively. The triangles are $\tilde{\mathbb{E}}[FD(k)]$ versus $\tilde{\mathbb{E}}[\widehat{FD}(k)]$ in 100 replications, and the dashed line is $x = y$. Right panels: the coverage frequencies of the 95% bootstrap confidence interval and confidence upper bound of $\mathbb{E}[FD(k)]$ (approximated by $\tilde{\mathbb{E}}[FD(k)]$), respectively, for k ranging from 50 to 70. The color shading represents different values of k.

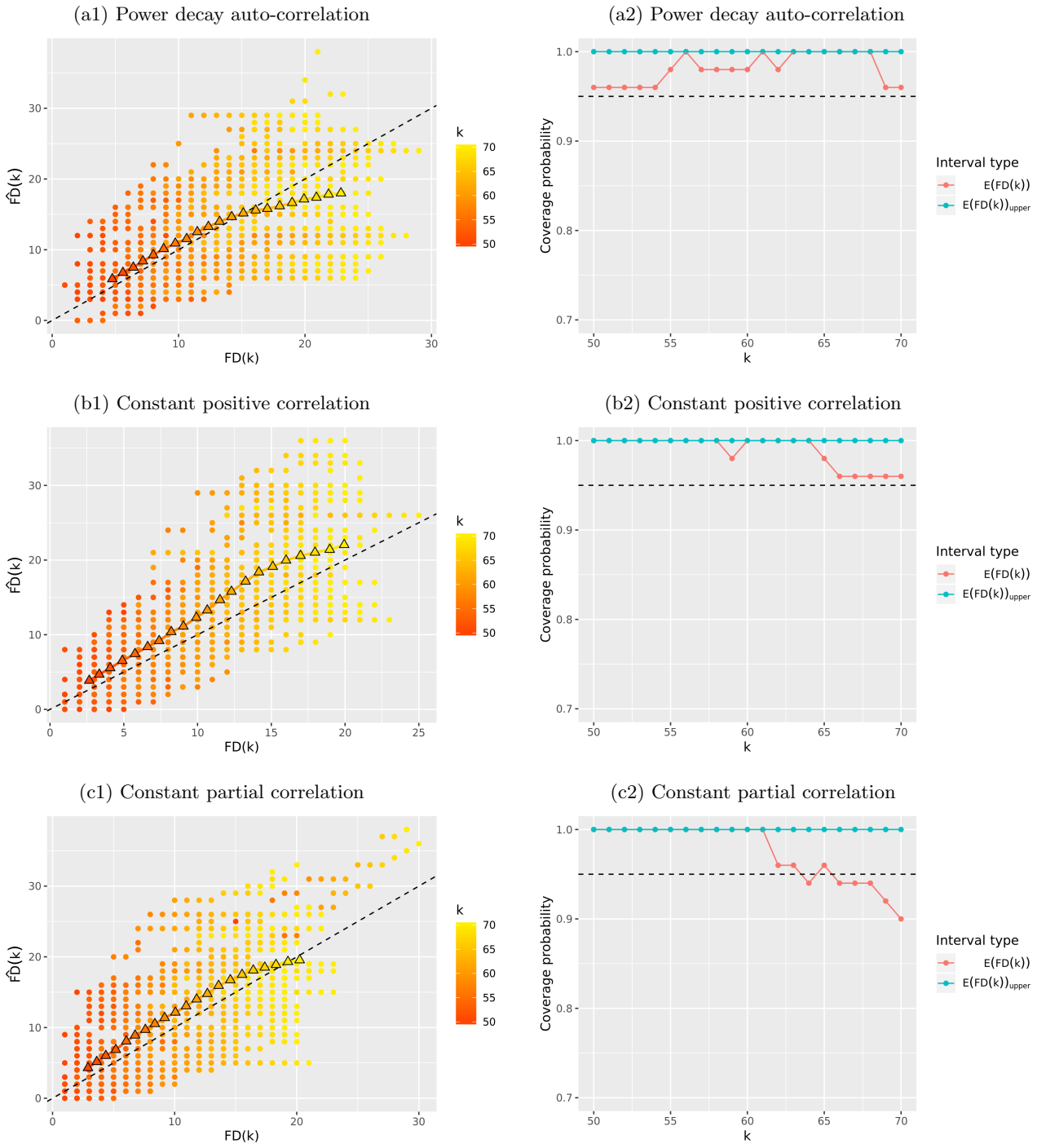

Figure 7: High dimensional scenario with (n=300, p=1000). Details and notations are the same as those in Figure 6.

k is smaller than 60. When k is larger than 60, the coverage probabilities drop occasionally below 0.95 by a small amount. This is expected since the variation of $\widehat{FD}(k)$ increases as k increases, which makes it more difficult to approximate its distribution.

Figure 7 shows the simulation results for high dimensional scenarios. For the auto-correlation structure, the scatter plot in Figure 7 (a1) shows that our proposed estimate tends to underestimate slightly when k is large. For the constant correlation structure and constant partial correlation structure (Figure 7 (b1, c1)), however, the proposed estimate slightly overestimate the number of false discoveries. Similar to the observation in the low dimensional case, the scatter points become more dispersed from the diagonal line as k increases. Figures 7 (a2,b2,c2) show the coverage of the two-sided confidence interval and confidence upper bound for $\mathbb{E}[FD(k)]$. The coverage probabilities of the confidence upper bound of $\mathbb{E}[FD(k)]$ were over 0.95 for all cases, and that for confidence interval are also mostly above 0.95 except for the constant partial correlation case with $k > 65$. This slight under-coverage appears to be a consequence of over-estimation of the number of FDs.

7 Discussion

We introduce the perturbation-based Gaussian Mirror method for controlling FDR in selecting variables for high-dimensional linear regression models. Intuitively, to test if a variable X_j is truly informative, the GM constructs a pair of variables mirroring each other at X_j. The scale of the mirror can be computed explicitly and easily as a function of the design matrix so as to guarantee that the distribution of the mirror test statistic is symmestric about zero under the null. With this construct, we show that asymptotically the FDR is controlled at any designated level. We have further proposed a way to assess the variance of the number of estimated FDs in a top-k list, providing extra information about the reliability of the reported FDP results. Through empirical studies we find that the GM is not very sensitive to the scale of the mirror, making the method more broadly applicable.

A distinctive advantage of the GM method is that it can be applied to any low or high-dimensional regression problems without requiring any distributional assumption or knowledge on the design matrix. It is thus ideal for analyzing the data arising from many application fields, such as population genetics, molecular and cellular biology, and biomedical researches. Furthermore, the GM method is constructed marginally for each covariate with twofold advantages: (i) unlike global constructions such as those in Model-X knockoffs, the mirror construction introduces only a small and controllable additive "disturbance", which not only alleviates possible high correlations among the original covariates, but also gains in power; (ii) the computation for the GM method can be easily parallelized, thus easily scalable to handle high dimensional large datasets, if advanced

computation infrastructures such as GPU are available. Unlike other marginal methods, such as the conditional randomization (Candes et al., 2018) relying on the conditional distribution of $[X_j \mid \mathbf{X}_{-j}]$, which is difficult or impossible to obtain in practice and highly expensive computationally, the construction of the GM is straightforward and nearly universal for both low and high-dimensional cases.

The GM design introduced in this article is a general framework based on the idea of marginally perturbing each variable by adding Gaussian noises. It is expected that such a construction can be generalized to handle more complex statistical and machine-learning models such as generalized linear models, index models, additive models, neural networks, etc. In linear models, we choose the scale of the mirror c_j to annihilate the partial correlation between the mirrored variables X_j^+ and X_j^- so that the mirror statistics is symmetric under the null. A nature gerneralization is to choose c_j to minimize a dependence measure of X_j^+ and X_j^- conditioned on the remaining variables. Our preliminary results show that a modified GM method works well for selecting important variables in neural network models. We will leave more detailed studies along this line to future research.

8 Appendix

Proof of Lemma 5. We start with the first statement. For any integer N, let

$$S_N = \sum_{j=1}^{N} \left(1(j \in \mathcal{S}_0, M_j \leq -t) - \mathbb{E}1(j \in \mathcal{S}_0, M_j \leq -t)\right).$$

According Chebyshev's inequality, for any $\epsilon > 0$,

$$P(|S_N| > N\epsilon) \leq \frac{Var(S_N)}{N^2\epsilon^2}.$$

Assumption 2 implies that

$$\begin{aligned}\mathrm{Var}(S_N) =& \sum_{j\in\mathcal{S}_0} \mathrm{Var}\big(1(M_j \leq -t)\big) + \sum_{1\leq j\neq k\leq N, j,k\in\mathcal{S}_0} \mathrm{Cov}\big(1(M_j \leq -t), 1(M_k \leq -t)\big) \\ \leq& N + CN^{\alpha} \leq CN^{\max(1,\alpha)}.\end{aligned}$$

Consequently, $P(|S_N| > N\epsilon) \leq CN^{\max(-1,\alpha-2)}$, and

$$P(|S_{N^2}| > N^2\epsilon) \leq CN^{\max(-2,2(\alpha-2))}.$$

Note that $\alpha < \frac{3}{2}$, implying that $2(\alpha-2) < -1$. Therefore,

$$\sum_{N=1}^{\infty} P(|S_{N^2}| > N^2\epsilon) < \infty.$$

According to Borel-Cantelli's lemma, we know that

$$\frac{S_{N^2}}{N^2} \to 0, a.s..$$

Next, consider $D_N = \max_{N^2 \le k < (N+1)^2} |S_k - S_{N^2}|$. Then

$$\mathbb{E}D_N^2 = \mathbb{E}\left(\max_{N^2 \le k < (N+1)^2} |S_k - S_{N^2}|\right)^2 \le \sum_{N^2 \le k < (N+1)^2} \mathbb{E}(S_k - S_{N^2})^2.$$

Based on Assumption 2, the $\mathbb{E}(S_k - S_{N^2})^2$ can be bounded as

$$\mathbb{E}(S_k - S_{N^2})^2 = \mathbb{E}\Big| \sum_{j=N^2+1}^{k} (1(j \in \mathcal{S}_0, M_j \le -t) - \mathbb{E}1(j \in \mathcal{S}_0, M_j \le -t))\Big|^2 \le C(k - N^2)^{\max(1,\alpha)}.$$

Combining the above two equations together, we have

$$\mathbb{E}D_N^2 \le \sum_{N^2 \le k < (N+1)^2} C(k - N^2)^{\max(1,\alpha)} \le C(2N+1)^{\max(2,\alpha+1)}.$$

Apply Chebyshev's inequality, we knows that

$$P(D_N > N^2\epsilon) \le \frac{\mathbb{E}D_N^2}{N^4\epsilon^2} \le CN^{max(-2,\alpha-3)}.$$

The fact that $\alpha - 3 < -\frac{3}{2}$ indicates that $\sum_N P(D_N > N^2\epsilon) < \infty$. According to Borel-Cantellis' lemma,

$$\frac{D_N}{N^2} \to 0, a.s..$$

In summary, for any integer p, let $N = \lfloor \sqrt{p} \rfloor$, then

$$\frac{|S_p|}{p} \le \frac{S_{N^2} + D_N}{N^2} \to 0 \quad a.s..$$

Note that $\frac{p_0}{p_0+p_1} \to \pi_0$ with $\pi_0 > 0$, therefore,

$$\frac{|S_p|}{p_0} \le \frac{S_{N^2} + D_N}{N^2} \to 0 \quad a.s..$$

Namely,

$$\frac{V(t) - \sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_0, M_j \le -t)}{p_0} \to 0 \quad a.s.. \tag{32}$$

According to the definition of $G_0(t)$, then

$$\frac{V(t)}{p_0} - G_0(t) \to 0 \quad a.s..$$

With similar argument, we can show that $\frac{U(t)}{p_1} - G_2(t) \to 0$, $\frac{W(t)}{p_1} - G_1(t) \to 0$ a.s.. For $R(t)$, similar argument can show that

$$R(t) \to \lim_p \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}1(M_j \geq t) \quad a.s..$$

Note that

$$\begin{aligned} \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}1(M_j \geq t) &= \frac{1}{p} \left[\sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_0, M_j \geq t) + \sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_1, M_j \geq t) \right] \\ = \quad \frac{1}{p} &\left[\sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_0, M_j \leq -t) + \sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_1, M_j \geq t) \right]. \end{aligned}$$

Therefore,

$$\lim_p \frac{1}{p} \sum_{j=1}^{p} \mathbb{E}1(M_j \geq t) = \pi_0 G_0(t) + \pi_1 G_1(t).$$

Next, we will show that $\frac{V(t)}{p_0}$ converges to $G_0(t)$ uniform almost surely. Note that $G_0(t)$ is continuous and bounded between 0 and 1. Therefore, for any $\epsilon > 0$, we can choose $0 = t_0 \leq t_1 \leq \cdots \leq t_k$ such that $|G_0(t_k) - G_0(t_{k-1})| < \epsilon, \forall k$. Assume that t falls in the interval (t_{k-1}, t_k), then $V(t_{k-1}) \leq V(t) \leq V(t_k)$. Therefore,

$$\sup_t |V(t)/p_0 - G_0(t)| \leq \epsilon + \sup_k |G_0(t_k) - G_0(t_{k-1})| \to 0 \quad a.s.$$

Similarly, one can show that $U(t)/p_1 - G_1(t)$ converges to 0 uniform almost surely, $W(t)/p_1 - G_1(t)$ converges to 0 uniform almost surely and $R(t)/p - \pi_0 G_0(t) + \pi_1 G_1(t)$ converges to zero uniform almost surely.

□

Proof of Theorem 6. Suppose that Assumption 1 holds. Using the l_1 convergence results in Van de Geer et al. (2014), we have

$$P(M_j > 0) = P(sign(\hat{\beta}_j^+) = sign(\hat{\beta}_j^-)) > 1 - \frac{2}{p}$$

for $j \in S_1$. By the Boole's inequality, we have that the event $\mathcal{E}$ holds with probability larger than $1 - \frac{2}{p}$. By (2), we have

$$\widehat{FD}(k) = \#\{M_j \leq -M_{(k)}\} = \#\{j : j \in \mathcal{S}_0, M_j < -M_{(k)}\} = V(M_{(k)}).$$

where $V(\cdot)$ is defined in Lemma 5. Also, $FD(k)$ can be written as

$$FD(k) = \#\{j \in S_0 : M_j \geq t\}.$$

The expection of $FD(k)$ is

$$\frac{1}{k}\mathbb{E}[FD(k)] = \sum_{j\in\mathcal{S}_0} \mathbb{E}1(M_j \geq M_{(k)}) = \sum_{j\in\mathcal{S}_0} \mathbb{E}1(M_j \leq -M_{(k)})$$

where the second equality holds by the symmetric property of M_j for $j \in S_0$. By equation (32) in the proof of Lemma 5, conditioned on $\mathcal{E}$, we have

$$\frac{V(t) - \sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_0, M_j \leq -t)}{k} \to 0 \quad a.s..$$

Since the event $\mathcal{E}$ holding with probability $1 - \frac{2}{p}$, we have

$$\frac{V(t) - \sum_{j=1}^{p} \mathbb{E}1(j \in \mathcal{S}_0, M_j \leq -t)}{k} \xrightarrow{p} 1.$$

□

Proof of Theorem 7 Assume that $\mathcal{E}$ holds, we have

$$\widehat{FD}(k) = \#\{M_j \leq -M_{(k)}\} = \#\{j : j \in \mathcal{S}_0, M_j < -M_{(k)}\}.$$

Thus we have that

$$\mathbb{E}[\widehat{FD}(k)] = \mathbb{E}1(j \in \mathcal{S}_0, M_j \leq -M_{(k)})) = \mathbb{E}1(j \in \mathcal{S}_0, M_j \geq M_{(k)})) = \mathbb{E}[FD(k)].$$

holds with probability $1 - \frac{2}{p}$. □

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

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

Barber, R. F. and E. J. Candès (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics 47*(5), 2504–2537.

Barber, R. F., E. J. Candès, and R. J. Samworth (2018). Robust inference with knockoffs. *To appear in Annals of Statistics arXiv:1801.03896*.

Bauchet, G., S. Grenier, N. Samson, J. Bonnet, L. Grivet, and M. Causse (2017). Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by genome wide association study. *Theoretical and Applied Genetics 130*(5), 875–889.

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

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics 41*(2), 802–837.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

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

Candès, E. J. and Y. Plan (2009). Near-ideal model selection by $l1$ minimization. *The Annals of Statistics 37*(5A), 2145–2177.

Chung, E. and J. P. Romano (2016). Multivariate and multiple permutation tests. *Journal of Econometrics 193*(1), 76–91.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika 62*(2), 441–444.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. pp. 569–593. Springer.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*, 191–203.

Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association 107*(499), 1019–1035.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics 44*(3), 907–927.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics 37*(1), 246–270.

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

Taylor, J. and R. Tibshirani (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics 46*(1), 41–61.

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

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics 7*, 1456–1490.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association 111*(514), 600–620.

Van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*(2), 614–645.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *The Annals of Statistics 37*(5A), 2178.

Yu, B. (2013). Stability. *Bernoulli 19*(4), 1484–1500.

Yu, B. and K. Kumbier (2019). Three principles of data science: predictability, computability, and stability (pcs). *arXiv preprint arXiv:1901.08152*.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics 36*(4), 1567–1594.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*(Nov), 2541–2563.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.