A Scale-free Approach for False Discovery Rate Control in Generalized Linear Models

Chenguang Dai*, Buyu Lin*, Xin Xing, and Jun S. Liu

Department of Statistics, Harvard University

July 3, 2020

Abstract

The generalized linear models (GLM) have been widely used in practice to model non-Gaussian response variables. When the number of explanatory features is relatively large, scientific researchers are of interest to perform controlled feature selection in order to simplify the downstream analysis. This paper introduces a new framework for feature selection in GLMs that can achieve false discovery rate (FDR) control in two asymptotic regimes. The key step is to construct a *mirror statistic* to measure the importance of each feature, which is based upon two (asymptotically) independent estimates of the corresponding true coefficient obtained via either the data-splitting method or the Gaussian mirror method. The FDR control is achieved by taking advantage of the mirror statistics property that, for any null feature, its sampling distribution is (asymptotically) symmetric about 0. In the moderate-dimensional setting in which the ratio between the dimension (number of features) p and the sample size n converges to a fixed value, i.e., $p/n \to \kappa \in (0, 1)$, we construct the mirror statistic based on the maximum likelihood estimation. In the high-dimensional setting where $p \gg n$, we use the debiased Lasso to build the mirror statistic. Compared to the Benjamini-Hochberg procedure, which crucially relies on the asymptotic normality of the Z statistic, the proposed methodology is scale free as it only hinges on the symmetric property, thus is expected to be more robust in finite-sample cases. Both simulation results and a real data application show that the proposed methods are capable of controlling the FDR, and are often more powerful than existing methods including the Benjamini-Hochberg procedure and the knockoff filter.

1 Introduction

The generalized linear model (GLM) is a useful tool for modeling non-Gaussian response variables such as categorical data and count data. In the current big data era, researchers are often capable of collecting a large number of explanatory features $X_1, \cdots, X_p$ for a given response variable y, in which the number of features p is potentially comparable to or larger than the sample size n. As the response variable y most likely only depends on a small subset of features, it is of primary interest to identify those relevant features in order to enhance the computability of the analysis as well as the interpretability of the results.

A desired feature selection procedure is expected to control the quality of the selection, which can be mathematically calibrated by the false discovery rate (FDR) (Benjamini and Hochberg, 1995) defined as:

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}], \quad \mathrm{FDP}^{1} = \frac{\#\{j : j \notin S_1,\ j \in \widehat{S}\}}{\#\{j \in \widehat{S}\}}, \tag{1}$$

*These authors contribute equally to this work.

[1] We assume FDP = 0 if $\#\{j \in \widehat{S}\} = 0$.

where S_1 denotes the index set of relevant features, $\widehat{S}$ denotes the index set of selected features, and FDP refers to the false discovery proportion. The expectation is taken with respect to the randomness both in the data and in the selection procedure if it is not deterministic. Existing FDR control methods that can be applied to GLMs include the Benjamini-Hochberg (BHq) procedure (Ma et al. (2020) specifically consider the logistic regression model) and the knockoff filter (Candes et al., 2018; Lu et al., 2018; Huang and Janson, 2019). A brief discussion on these existing methods as well as their comparisons to our proposed strategy are given in Section 4.3.

In this paper, we develop new methodologies for exercising controlled feature selection in GLMs based upon the recently developed FDR control framework in Xing et al. (2019) and Dai et al. (2020). Under the guiding principle of data perturbation, we construct a mirror statistic for each feature to measure its relative importance, based upon two estimates of the corresponding true coefficient obtained via either the Gaussian mirror method (Xing et al., 2019) or the data-splitting method (Dai et al., 2020). After choosing a proper data-dependent cutoff, we select the features with mirror statistics larger than the cutoff. The FDR control is achieved by exploiting the symmetric property of the mirror statistic associated with any null feature. Our FDR control framework enjoys a scale-free property in the sense that any constant rescaling of all the mirror statistics does not change the selection result.

We consider two asymptotic regimes for GLMs in this paper. The moderate-dimensional setting concerns the regime where the ratio $p/n \to \kappa \in (0, 1)$. We base the construction of the mirror statistics on the maximal likelihood estimator (MLE) of the true coefficient vector and show that both the Gaussian mirror and the data-splitting methods achieve an asymptotic FDR control under mild conditions. We note that the classical asymptotic result for the MLE breaks down in this regime, in the sense that the asymptotic normality characterization of the MLE involves two additional bias and variance scaling factors (Sur and Candès, 2019). In consequence, BHq faces the challenge of estimating the two scaling factors in order to obtain asymptotically valid p-values, which remains an open problem for GLMs except for the logistic/probit regression model. In contrast, our FDR control framework is scale-free and does not require the knowledge of the aforementioned scaling factors, thus can be easily and validly applied to all GLMs.

The high-dimensional setting concerns the regime of $p \gg n$. We restrict ourselves to the data-splitting method for the consideration of computational feasibility. We construct the mirror statistics using the debiased Lasso (Van de Geer et al., 2014), and theoretically justify our approach by showing the desired FDR control property under proper sparsity and regularity conditions. Ma et al. (2020) introduced a BHq procedure for the logistic regression model, which relies on the asymptotic normality of the debiased Lasso estimator in order to obtain asymptotically valid p-values. In contrast, for the purpose of controlling the FDR, the data-splitting method only requires the asymptotic symmetric property of the debiased Lasso estimator regardless of its scale, thus we expect it to be more robust in finite-sample cases.

The rest of the paper is structured as follows. Section 2 introduces our FDR control framework as well as two basic methods for constructing the mirror statistics: Gaussian mirror and data splitting. Section 3 concerns the GLMs in the moderate-dimensional setting. We specify the construction of the mirror statistics using the MLE, and establish the desired FDR control property for both the Gaussian mirror and the data-splitting methods. Section 4 concerns the GLMs in the high-dimensional setting. By incorporating the debiasing approach, we show that the data-splitting method enjoys an asymptotic FDR control. Sections 5.1 and 5.2 demonstrate the competitive performances of our proposed methods through simulation studies on popular GLMs including the logistic regression model and the negative binomial regression model. Section 5.3 applies the data-splitting method to a single-cell RNA sequencing data for the purpose of selecting relevant genes with respect to the glucocorticoid response. Section 6 concludes with a few final remarks. The proofs as well as additional numerical results are given in the Appendix.

2 FDR Control via Mirror Statistics

For a given response variable y, we consider a set of p candidate features $\{X_1, \ldots, X_p\}$. Let $X_{n\times p}$ be the design matrix, in which each row, denoted as $x_i^\intercal$ for $i \in [n]$, is an independent realization of these features. Let $y = (y_1, \cdots, y_n)^\intercal$ be the associated response vector. We assume that the response variable y only depends on a subset of features, and the corresponding index set is denoted as S_1. Let $p_1 = |S_1|$ and $p_0 = p - p_1$. We refer the feature X_j as relevant (non-null) if $j \in S_1$; otherwise we call it a null feature. The index set of null features are denoted as S_0. The goal is to identify as many relevant features as possible with the FDR under control. Throughout, we denote the power of a selection procedure as the expected proportion of the successfully identified relevant features, i.e.,

$$\text{Power} = \mathbb{E}\big[\#\{j : j \in S_1, j \in \widehat{S}\}/p_1\big], \tag{2}$$

in which $\widehat{S}$ denotes the index set of selected features.

The FDR control framework we consider here requires constructing a mirror statistic M_j for each feature X_j, which satisfies the following two properties.

(A1) A feature with a larger mirror statistic is more likely to be a relevant feature.

(A2) The sampling distribution of the mirror statistic associated with any null feature is (asymptotically) symmetric about 0.

Property (A1) enables us to rank the importance of features by their associated mirror statistics. Given the FDR control level q, it remains to choose a proper cutoff τ_q, and select the set of features $\{j : M_j > \tau_q\}$. For any cutoff $t > 0$, Property (A2) suggests an approximate upper bound on the number of false positives,

$$\text{FDP}(t) = \frac{\#\{j : j \notin S_1, M_j > t\}}{\#\{j : M_j > t\}} \lesssim \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\}}, \tag{3}$$

which leads to the following FDR control framework.

Algorithm 1 The FDR control framework.

1. Calculate the mirror statistic M_j for $j \in [p]$.
2. Given a designated FDR level $q \in (0, 1)$, calculate the cutoff τ_q as:
$$\tau_q = \inf\left\{t > 0 : \widehat{\text{FDP}}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\}} \le q\right\}.$$
3. Set $\widehat{S} = \{j : M_j > \tau_q\}$.

A general recipe for constructing the mirror statistics in the regression setting is as follows. For $j \in [p]$, we first obtain two estimates, $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$, of the true coefficient $\beta_j^\star$. In order for the resulting mirror statistics to have comparable variances, the two estimates generally require further standardization by the corresponding (asymptotic) standard deviations. Denote $T_j^{(1)}$ and $T_j^{(2)}$ as the normalized estimates, which should satisfy the following conditions.

Condition 2.1.

- (Independence) The two regression coefficients are (asymptotically) independent.
- (Symmetry) For any null feature $j \in S_0$, the sampling distribution of either of the two regression coefficients is (asymptotically) symmetric about 0.

The mirror statistic M_j then takes a general form of

$$M_j = \text{sign}\big(T_j^{(1)} T_j^{(2)}\big) f\big(|T_j^{(1)}|, |T_j^{(2)}|\big), \tag{4}$$

in which $f(u, v)$ is a user-specified bivariate function satisfying the following conditions:

Condition 2.2. $f(u, v)$ is non-negative, symmetric and monotonically increasing in u and v.

Condition 2.1 and 2.2 together implies Property (A1) and (A2). For a relevant feature $j \in S_1$, the two regression coefficients $T_j^{(1)}$ and $T_j^{(2)}$ tend to be large (in the absolute value) and have the same sign if the estimation procedures are reasonably accurate. Since $f(u, v)$ is monotonically increasing in both u and v, the mirror statistic M_j is likely to be positive and relatively large, which implies Property (A1). On the other hand, for a null feature $j \in S_0$, Property (A2) holds given Condition 2.1 because $T_j^{(1)}$ and $T_j^{(2)}$ are (asymptotically) independent, and one of them is (asymptotically) symmetric about 0.

Three possible choices of $f(u, v)$ are

$$f(u, v) = 2\min(u, v), \quad f(u, v) = uv, \quad f(u, v) = u + v. \tag{5}$$

The first choice is equal to the mirror statistic proposed in Xing et al. (2019), and the third choice is nearly optimal as shown by the following proposition.

Proposition 2.1. *Suppose the set of normalized regression coefficients $\{T_j\}_{j\in[p]}$ are asymptotically independent across the feature index j. For each null feature $j \in S_0$, we assume that T_j asymptotically follows $N(0, 1)$, whereas for each relevant feature $j \in S_1$, we assume that T_j asymptotically follows $N(\omega, 1)$ with $\omega > 0$. In addition, we assume that p_0/p converges to a fixed constant in $(0, 1)$. Then $f(u, v) = u + v$ is the nearly optimal choice satisfying Condition 2.2 that yields the highest power, when ω is sufficiently large.*

Remark 2.1. The proof of Proposition 2.1 might be of interest on its own. We essentially rephrase the FDR control problem under the hypothesis testing framework, and prove the optimality using the Neymann-Pearson Lemma. The form $f(u, v) = u + v$ is derived based on the rejection rule of the corresponding likelihood ratio test. In practice, however, we found that the three choices listed in (5) have no significant differences in most cases, and the second one $f(u, v) = uv$, which will be used in our simulation studies, sometimes can even perform slightly better.

The following subsections review two recently proposed methods, Gaussian mirror (Xing et al., 2019) and data splitting (Dai et al., 2020), for constructing the two regression coefficients $T_j^{(1)}$ and $T_j^{(2)}$ that satisfy Condition 2.1.

2.1 Gaussian mirror

For an easy illustration, we restrict ourselves to low-dimensional ($n > p$) linear models. For high-dimensional settings, we refer the readers to Xing et al. (2019). The key idea of Gaussian mirror is to create a pair of perturbed mirror features,

$$X_j^+ = X_j + c_j Z_j, \quad X_j^- = X_j - c_j Z_j, \tag{6}$$

in which c_j is an adjustable scalar and Z_j follows $N(0, 1)$ independently. The linear model can then be equivalently recasted in the following way,

$$y = \frac{\beta_j^\star}{2} X_j^+ + \frac{\beta_j^\star}{2} X_j^- + X_{-j}\beta_{-j}^\star + \epsilon. \tag{7}$$

As we are in the low-dimensional setting, we obtain $\widehat{\beta}^+$ and $\widehat{\beta}^-$, as well as the normalized estimates T_j^+ and T_j^-, via the ordinary least squares (OLS). For any null feature $j \in S_0$, both T_j^+ and T_j^- follow a t distribution centering at zero, thus the symmetric requirement in Condition 2.1 is fulfilled. Furthermore, we can properly set c_j as follows so that T_j^+ and T_j^- are asymptotically independent,

$$c_j = ||P_{-j}^{\perp} X_j||/||P_{-j}^{\perp} Z_j||, \tag{8}$$

where $P_{-j}^{\perp}$ is the projection matrix onto the orthogonal complement of the column space spanned by X_{-j}.

2.2 Data splitting

A simple way to obtain two independent regression coefficients is data splitting. More precisely, we randomly split the data into two halves, $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$, and estimate $\widehat{\beta}_j^{(1)}$ and $\widehat{\beta}_j^{(2)}$, as well as the normalized versions $T_j^{(1)}$ and $T_j^{(2)}$, on each part of the data. Without loss of generality, we assume that the sample sizes of the two parts of the data are the same. The independence between the two estimates is naturally implied by data splitting. The symmetric requirement in Condition 2.1 can be satisfied if, for any null feature, either of the estimates is (asymptotically) normal and centered at 0. As we will see, desirable estimates can be constructed for GLMs under proper conditions.

The potential power loss is a major concern of using data splitting. Dai et al. (2020) introduced a multiple data-splitting (MDS) method to remedy this issue, which also helps in stabilizing the selection result. The idea is to obtain multiple selection results via repeated data splits, and rank the importance of features by their inclusion rates defined as below,

$$\widehat{I}_j = \frac{1}{m} \sum_{k=1}^{m} \frac{\mathbb{1}(j \in \widehat{S}^{(k)})}{|\widehat{S}^{(k)}| \vee 1}, \tag{9}$$

in which m is the total number of data splits, and $\widehat{S}^{(k)}$ is the index set of the selected features in the k-th data split. We select the features with inclusion rates larger than a properly chosen cutoff so as to maintain the FDR control. Dai et al. (2020) also show that, for the independent Gaussian means problem, MDS can recover the full information in the sense that when $m \to \infty$, the inclusion rates provide the same ranking of features as the p-values calculated based on the full data. We outline the MDS procedure in Algorithm 2, which can be applied on top of the single data-splitting methods designed for GLMs (see Sections 3 and 4).

Algorithm 2 Aggregating selection results from multiple data splits.

1. Sort the features with respect to their inclusion rates in the multiple selections. Denote the sorted inclusion rates as $0 \le \widehat{I}_{(1)} \le \widehat{I}_{(2)} \le \cdots \le \widehat{I}_{(p)}$.
2. Given a designated FDR level $q \in (0, 1)$, find the largest $\ell \in [p]$ such that $\widehat{I}_{(1)} + \cdots + \widehat{I}_{(\ell)} \le q$.
3. Set $\widehat{S} = \{j : \widehat{I}_j > \widehat{I}_{(\ell)}\}$.

3 Generalized Linear Models in Moderate Dimensions

We consider the following generalized linear model (GLM) with a canonical link function ρ:

$$p(y|X,\beta^\star) = \prod_{i=1}^{n} c(y_i) \exp\left(y_i x_i^\intercal \beta^\star - \rho(x_i^\intercal \beta^\star)\right), \tag{10}$$

where $\beta^\star$ denotes the true coefficient vector. In the moderate-dimensional setting, we assume that $p/n \to \kappa \in (0,1)$. Note that $\kappa = 0$ corresponds to the classical setting with fixed p. We consider a random design setting, in which we assume that the x_i's are i.i.d observations from $N(0,\Sigma)$. We assume that the signal strength also has an asymptotic limit, i.e., $\mathrm{Var}(x_i^\intercal \beta^\star) \to \gamma^2$.

Let $\rho(X\beta) = (\rho(x_1^\intercal\beta), \cdots, \rho(x_n^\intercal\beta))^\intercal$. We base the mirror statistics on the MLE,

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} 1^\intercal \rho(X\beta) - \frac{1}{n} y^\intercal X\beta \right\}. \tag{11}$$

The MLE can behave very differently in the moderate-dimensional setting compared to the classical setting with fixed p. First of all, the MLE may not exist. For instance, for the logistic regression model, the MLE does not exist if the two classes become well separated. More precisely, Candès and Sur (2020) derived the phase transition curve for the existence of MLE, parametrized in terms of κ and γ. In this section, we restrict ourselves to the regime where the MLE exists, otherwise we can consider the debiased regularization method detailed in Section 4. Second, when the MLE exists, it is still asymptotically biased and rescaled. Based upon the results of Zhao et al. (2020) and Salehi et al. (2019), we obtain the following characterization of the asymptotic distribution of the MLE.

Proposition 3.1. *Consider a GLM defined in (10) in the moderate-dimensional setting as specified above. Assume that the MLE exists asymptotically. Then, for any $j \in [p]$ with $\sqrt{n}\tau_j\beta_j^\star = O(1)$, we have*

$$\frac{\sqrt{n}(\widehat{\beta}_j - \alpha_\star \beta_j^\star)}{\sigma_\star/\tau_j} \xrightarrow{d} N(0,1), \tag{12}$$

where $\tau_j^2 = 1/\Theta_{jj}$ with $\Theta = \Sigma^{-1}$, and $\alpha_\star, \sigma_\star$ are two universal constants depending on the model ρ, the true coefficient vector $\beta^\star$, the signal strength γ and the ratio of dimension to sample size κ.

Remark 3.1. The proof of Proposition 3.1 relies on the stochastic representation of the MLE in Zhao et al. (2020) and the Convex Gaussian Min-max (CGMT) Theorem. The constant pair $(\alpha_\star, \sigma_\star^2)$ is the limit of (α_n, σ_n^2), where

$$\alpha_n = \frac{\langle \widehat{\beta}, \beta^\star \rangle}{||\widehat{\beta}||^2}, \qquad \sigma_n^2 = ||P_{\beta^\star}^\perp \widehat{\beta}||. \tag{13}$$

$P_{\beta^\star}^\perp$ is the projection matrix onto the orthogonal complement of $\beta^\star$. The convergence of (α_n, σ_n^2) follows from a routine application of CGMT by transforming the primary optimization problem (PO) into an easy-to-analyze auxiliary optimization problem (AO). More details can be found in the proof of Lemma A.1.

Remark 3.2. Note that τ_j^2 is simply the conditional variance $\mathrm{Var}(X_j|X_{-j})$, and thus can be estimated either using the inverse of the sample covariance matrix, i.e., $\widehat{\tau}_j^2 = 1/(X^\intercal X/n)_{jj}^{-1}$, or via a node-wise regression approach, that is to regress X_j onto X_{-j} and obtain the residual sum of squares RSS_j. An unbiased estimator of τ_j^2 is then $\widehat{\tau}_j^2 = \mathrm{RSS}_j/(n-p+1)$.

Proposition 3.1 implies that the classical asymptotic normality based on the Fisher information does not apply to the MLE in the moderate-dimensional setting. In addition, in order to obtain asymptotically valid p-values, it requires to estimate the bias and variance scaling factors $\alpha_\star$ and $\sigma_\star$.

This is in general a challenging task, as it requires to first estimate the signal strength γ. Sur and Candès (2019) proposed the *ProbFrontier* method to estimate γ using the phase transition curve that calibrates the existence of the MLE. However, to the best of our knowledge, there is no unified approach to derive the curve for a general GLM and the existing literature only covers the results for the logistic/probit regression model. Therefore, it remains challenging to apply BHq moving beyond these two models.

Even for the logistic/probit regression model, we empirically observe that the asymptotic normality characterized in Proposition 3.1 might kick in slowly, so that the resulting p-value of the null feature appears non-uniform in finite-sample cases. In contrast, the asymptotic symmetry required by our FDR control framework can kick in much faster compared to the asymptotic normality. Numerical comparisons can be found in Section 5.1.1.

3.1 FDR control via data splitting

Contrast to BHq, the data-splitting method outlined in Algorithm 3 is free of estimating the scaling factors $(\alpha_\star, \sigma_\star)$, thus is applicable for all GLMs in the moderate-dimensional setting. According to the asymptotic characterization in Equation (12), we normalize the two independent MLEs $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ as below,

$$T_j^{(1)} = \widehat{\tau}_j^{(1)} \widehat{\beta}_j^{(1)}, \quad T_j^{(2)} = \widehat{\tau}_j^{(2)} \widehat{\beta}_j^{(2)}. \tag{14}$$

Here $\widehat{\tau}_j^{(1)}$ and $\widehat{\tau}_j^{(2)}$ are two independent estimates of the conditional variance $\mathrm{Var}(X_j|X_{-j})$ obtained via either the node-wise regression or the inverse of the sample covaraince matrix. Although the asymptotic standard deviation of the MLE is $\sigma_\star/\tau_j$, we can safely drop the constant $\sigma_\star$, because our FDR control framework outlined in Algorithm 1 maintains the same selection result under an arbitrary rescaling of all the mirror statistics. Furthermore, the data-splitting method does not require the knowledge of the bias scaling factor $\alpha_\star$ either. For any null feature $j \in S_0$, since $\alpha_\star \beta_j^\star = 0$, the symmetric requirement in Condition 2.1 is asymptotically fulfilled according to Proposition 3.1.

Algorithm 3 The data-splitting method for GLMs in the moderate-dimensional setting.

1. Split the data set into two equal-sized halves $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$.
2. For $j \in [p]$, regress $X_j^{(1)}$ onto $X_{-j}^{(1)}$, and regress $X_j^{(2)}$ onto $X_{-j}^{(2)}$. Let

$$\widehat{\tau}_j^{2(1)} = \frac{\mathrm{RSS}_j^{(1)}}{n/2 - p + 1}, \quad \widehat{\tau}_j^{2(2)} = \frac{\mathrm{RSS}_j^{(2)}}{n/2 - p + 1},$$

   in which RSS_j is the residual sum of squares.
3. Find the MLEs $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ on each part of the data. For $j \in [p]$, calculate the mirror statistic M_j based on $T_j^{(1)}$ and $T_j^{(2)}$ defined in Equation (14).
4. Select the features using Algorithm 1.

We require the following assumptions to theoretically justify our approach. Note that we impose no assumptions on either the sparsity level or the signal magnitude of the true coefficient vector $\beta^\star$.

Assumption 3.1.

(1) There exists a constant $C > 0$, such that $1/C \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le C$.

(2) The required number of null features $p_0 \to \infty$ as $n, p \to \infty$.

Remark 3.3. Assumption 3.1 (1) also appears in Zhao et al. (2020), in which $\sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma)$ refer to the minimum and maximum of the eigenvalues of the covariance matrix Σ. Assumption 3.1 (2) is straightforward because otherwise the asymptotic FDR control problem becomes trivial: we can simply select all features and the corresponding FDR converges to 0.

Under Assumption 3.1, the following proposition shows that the data-splitting method achieves an asymptotic FDR control for a GLM in the moderate-dimensional setting.

Proposition 3.2. *Consider a GLM defined in (10) in the moderate-dimensional setting. For any given FDR control level $q \in (0,1)$, we assume that the pointwise limit $FDP^{\infty}(t)$ of $FDP(t)$ exists for all $t > 0$, and there is a $t_q > 0$ such that $FDP^{\infty}(t_q) \le q$. Then, under Assumption 3.1, we have*

$$\limsup_{n,p\to\infty} \mathbb{E}\left[\frac{\#\{j : j \in S_0, j \in \widehat{S}_{\tau_q}\}}{\#\{j : j \in \widehat{S}_{\tau_q}\}}\right] \le q.$$

using the data-splitting method outlined in Algorithm 3.

Remark 3.4. The assumption on the existence of a desirable t_q is necessary, as it implies that the asymptotic FDR control is achievable by selecting a proper cutoff for the mirror statistics.

3.2 FDR control via Gaussian mirror

The data-splitting method is only applicable in the regime $\kappa \in (0, 1/2]$, i.e., $n \ge 2p$; otherwise the MLE does not exist after data splitting. In fact, for the logistic regression model, even in the regime $\kappa \in (0, 1/2]$, it is still possible, if the signal strength γ is sufficiently large, that the MLE exists on the full data but not on a half of the data (e.g., see Figure 6(a) in Sur and Candès (2019)). To overcome this issue, we consider the Gaussian mirror method, which extends the applicability to $\kappa \in (0, 1)$ as long as the MLE exists on the full data.

As discussed in Section 2.1, we fit a GLM using the response vector y and the augmented set of features (X_{-j}, X_j^+, X_j^-), to find the MLEs, $\widehat{\beta}^+$ and $\widehat{\beta}^-$, associated with the pair of perturbed mirror features (X_j^+, X_j^-) defined in Equation (6). Let Σ_{aug} be the covariance matrix of the augmented set of features (X_{-j}, X_j^+, X_j^-), and let $\Theta_{\text{aug}} = \Sigma_{\text{aug}}^{-1}$. We have the following asymptotic characterization.

Proposition 3.3. *For any $j \in [p]$, consider fitting a GLM using the response vector y and the augmented set of features (X_{-j}, X_j^+, X_j^-) defined in Equation (6). Then, the asymptotic distribution of the MLE $(\widehat{\beta}_j^+, \widehat{\beta}_j^-)$ is:*

$$\frac{\sqrt{n}}{\sigma_\star}\left(\begin{pmatrix}\widehat{\beta}_j^+\\ \widehat{\beta}_j^-\end{pmatrix} - \frac{\alpha_\star}{2}\begin{pmatrix}\beta_j^\star\\ \beta_j^\star\end{pmatrix}\right) \xrightarrow{d} N\big(0, \Theta^*\big), \tag{15}$$

in which Θ^* is the 2×2 submatrix at the right bottom of Θ_{aug} corresponding to (X_j^+, X_j^-), and $\alpha_\star, \sigma_\star$ are defined as in Proposition 3.1.

According to Proposition 3.3, we can choose a proper scalar c_j so that $\Theta^*_{12} = 0$. This implies that the MLEs $\widehat{\beta}_j^+$ and $\widehat{\beta}_j^-$ are asymptotically independent. In practice, we plug in the sample covariance matrix $X^\intercal X/n$ to estimate the population covariance matrix Σ, leading to the scalar c_j in the form of Equation (8). We note that $\Theta^*_{12} = 0$ also implies the asymptotic independence between X_j^+ and X_j^- conditioning on X_{-j}. Therefore, for our specific choice of c_j, the inverse of the asymptotic variances of $\widehat{\beta}_j^+$ and $\widehat{\beta}_j^-$ are

$$\begin{aligned} 1/\Theta^*_{11} &= \mathrm{Var}(X_j^+ \mid X_j^-, X_{-j}) \approx \mathrm{Var}(X_j^+ \mid X_{-j}) = \tau_j^2 + c_j^2,\\ 1/\Theta^*_{22} &= \mathrm{Var}(X_j^- \mid X_j^+, X_{-j}) \approx \mathrm{Var}(X_j^- \mid X_{-j}) = \tau_j^2 + c_j^2. \end{aligned} \tag{16}$$

We thus normalize $\widehat{\beta}_j^+$ and $\widehat{\beta}_j^-$ as

$$T_j^+ = (\widehat{\tau}_j^2 + c_j^2)^{1/2}\widehat{\beta}_j^+, \quad T_j^- = (\widehat{\tau}_j^2 + c_j^2)^{1/2}\widehat{\beta}_j^-, \tag{17}$$

with $\widehat{\tau}_j^2$ calculated according to Remark 3.2. We summarize the Gaussian mirror method in Algorithm 4.

Algorithm 4 The Gaussian mirror method for GLMs in the moderate-dimensional setting.

1. For $j \in [p]$, calculate the mirror statistic M_j as follows.
   (a) Simulate Z_j from $N(0, I_n)$.
   (b) Calculate the scaling factor c_j according to Equation (8).
   (c) Fit a GLM using y and (X_{-j}, X_j^+, X_j^-) to find the MLEs, $\widehat{\beta}_j^+$ and $\widehat{\beta}_j^-$.
   (d) Estimate $\widehat{\tau}_j^2$ according to Remark 3.2.
   (e) Calculate the mirror statistic M_j based on T_j^+ and T_j^- defined in Equation (17).
2. Select the features using Algorithm 1.

Remark 3.5. The Gaussian mirror method is computationally more intensive compared to the data-splitting method. The former requires fitting the GLM p times, whereas the latter only requires fitting the GLM two times.

Under Assumption 3.1, the following proposition shows that the Gaussian mirror method achieves an asymptotic FDR control for a GLM in the moderate-dimensional setting.

Proposition 3.4. *Consider a GLM defined in (10) in the moderate-dimensional setting. For any given FDR control level $q \in (0,1)$, we assume that the pointwise limit $FDP^{\infty}(t)$ of $FDP(t)$ exists for all $t > 0$, and there is a $t_q > 0$ such that $FDP^{\infty}(t_q) \le q$. Then, under Assumption 3.1, we have*

$$\limsup_{n,p\to\infty} \mathbb{E}\left[\frac{\#\{j : j \in S_0, j \in \widehat{S}_{\tau_q}\}}{\#\{j : j \in \widehat{S}_{\tau_q}\} \vee 1}\right] \le q,$$

using the Gaussian mirror method outlined in Algorithm 4.

4 Generalized Linear Models in High Dimensions

In this section, we consider the high-dimensional setting ($p \gg n$), in which we base the mirror statistics on the regularized estimator instead of the MLE. To better illustrate the idea, we first investigate the high-dimensional linear model, upon which we extend the discussion to GLMs. In addition, considering the computational feasibility, we focus on the data-splitting method.

4.1 Linear models

4.1.1 Construction of the mirror statistics

Assume the true data generating process is a linear model $y = X\beta^\star + \epsilon$ where ϵ follows $N(0, \sigma^2 I_n)$. We consider the random design setting, where x_i's are i.i.d. random vectors with population covariance matrix Σ. Without loss of generality, we assume $\Sigma_{jj} = 1$ for $j \in [p]$. In the high-dimensional

setting, the mirror statistics are built upon the Lasso estimator (Tibshirani, 1996) defined as below,

$$\widehat{\beta}(y, X; \lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_1 \right\}. \tag{18}$$

The Lasso estimator of the null feature is generally biased except for cases with an orthogonal design matrix. In order to asymptotically remove the bias and symmetrize the Lasso estimator, we employ the debiasing approach introduced in Javanmard and Montanari (2014), Zhang and Zhang (2014), and Van de Geer et al. (2014). The debiased Lasso estimator $\widehat{\beta}^d$ takes the following simple form,

$$\widehat{\beta}^d = \widehat{\beta} + \frac{1}{n} D X^\intercal (y - X\widehat{\beta}), \tag{19}$$

in which D is a decorrelating matrix. Plugging in $y = X\beta^\star + \epsilon$, we obtain the following decomposition,

$$\sqrt{n}(\widehat{\beta}^d - \beta^\star) = Z + \Delta, \quad Z|X \sim N(0, \sigma^2 D\widehat{\Sigma}D^\intercal), \quad \Delta = \sqrt{n}(D\widehat{\Sigma} - I)(\beta^\star - \widehat{\beta}), \tag{20}$$

where $\widehat{\Sigma} = (X^\intercal X)/n$ is the sample covariance matrix. Let $\Lambda = D\widehat{\Sigma}D^\top$. The two terms Δ and $\sigma^2\Lambda$ calibrate the asymptotic bias and variance of the debiased Lasso estimator, respectively.

Various proposals of the decorrelating matrix D have been documented in the literature. Javanmard and Montanari (2014) proposed an optimization approach in order to simultaneously minimize the bias term Δ and the variance term Λ. In this paper, we follow the approach used in Javanmard and Montanari (2013) and Zhang and Zhang (2014), and set $D = \widehat{\Theta}$ as an estimator of the precision matrix $\Theta = \Sigma^{-1}$. One natural way to construct $\widehat{\Theta}$ is via regularized node-wise regression as detailed in Algorithm 5, which is based on the fact that $\Theta_{j,-j}$ corresponds to the coefficients of the best linear predictor of X_j using X_{-j}.

Algorithm 5 Construction of the decorrelating matrix $\widehat{\Theta}$.

1. Node-wise Lasso regression. For $j \in [p]$, let

$$\begin{aligned} \text{(Linear model)} \quad \widehat{\gamma}_j &= \underset{j \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} ||X_j - X_{-j}\gamma||_2^2 + \lambda_j ||\gamma||_1 \right\}; \\ \text{(GLM)} \quad \widehat{\gamma}_j &= \underset{j \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} ||X_{\widehat{\beta},j} - X_{\widehat{\beta},-j}\gamma||_2^2 + \lambda_j ||\gamma||_1 \right\}. \end{aligned} \tag{21}$$

2. Define $\widehat{C}$ with $\widehat{C}_{j,j} = 1$, and $\widehat{C}_{j,k} = -\widehat{\gamma}_{j,k}$ for $k \neq j$, where $\widehat{\gamma}_{j,k}$ is the k-th entry of $\widehat{\gamma}_j$.

3. Let $\widehat{\Theta} = \widehat{G}^{-2}\widehat{C}$, in which $\widehat{G} = \text{diag}(\widehat{\tau}_1^2, \cdots, \widehat{\tau}_p^2)$ with

$$\begin{aligned} \text{(Linear model)} \quad \widehat{\tau}_j^2 &= (X_j - X_{-j}\widehat{\gamma}_j)^\intercal X_j / n; \\ \text{(GLM)} \quad \widehat{\tau}_j^2 &= (X_{\widehat{\beta},j} - X_{\widehat{\beta},-j}\widehat{\gamma}_j)^\intercal X_j / n. \end{aligned} \tag{22}$$

Under proper conditions, the bias term Δ vanishes asymptotically. Thus, the symmetry requirement in Condition 2.1 is satisfied since $\sqrt{n}\widehat{\beta}_j^d$ asymptotically follows $N(0, \sigma^2\Lambda_{jj})$ for $j \in S_0$. After the normalization by the asymptotic variance, we obtain the normalized debiased Lasso estimator:

$$T_j = \widehat{\beta}_j^d / \widehat{\sigma}_j \quad \text{with} \quad \widehat{\sigma}_j^2 = \Lambda_{jj} = \left(\widehat{\Theta}\widehat{\Sigma}\widehat{\Theta}^\top\right)_{jj}, \quad \text{for } j \in [p]. \tag{23}$$

Remark 4.1. As discussed in Section 3.1, we can safely drop the constants $\sqrt{n}$ and σ in the construction of the mirror statistics. Thus the data-splitting method is free of estimating the noise level σ. This scale-free property potentially makes our approach more appealing compared to BHq (Javanmard and Javadi, 2019), which generally plugs in a consistent estimator of σ (such as the one estimated by the scaled Lasso (Sun and Zhang, 2012)) in order to obtain asymptotically valid p-values. Empirically, we observe that when the features are moderately correlated, the scaled Lasso tends to over estimate the true noise level. In consequence, the asymptotic p-values of the relevant features are right-skewed, leading to a power loss. Numerical comparisons can be found in Section 5.2.1.

The data-splitting method then proceeds by first randomly splitting the data into two halves, $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$, and then computing the two independent debiased Lasso estimates, $\widehat{\beta}^{(1,d)}$ and $\widehat{\beta}^{(2,d)}$ following Equation (19), where $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ are solutions to the optimization problem (18), and $\widehat{\Theta}^{(1)}$ and $\widehat{\Theta}^{(2)}$ are computed by Algorithm 5. The normalized estimators $T^{(1)}$ and $T^{(2)}$ follow Equation (23), and the mirror statistic M_j is constructed based on Equation (4) using a user-specified function $f(u, v)$ satisfying Condition 2.2. A summary of the data-splitting method is given in Algorithm 6.

Remark 4.2. Xing et al. (2019) and Dai et al. (2020) also consider the high-dimensional linear models based on the same FDR control framework described in Section 2. Xing et al. (2019) proposed to symmetrize the Lasso estimator via the post-selection procedure (Lee et al., 2016), while Dai et al. (2020) introduced a Lasso + OLS procedure, in which Lasso is first applied to one half of the data, and then OLS is applied to the other half of the data using the subset of features selected by Lasso. The symmetry requirement in Condition 2.1 is satisfied as long as all relevant features are selected by Lasso in the first step. However, this may not be justified for GLMs. In contrast, as shown in Section 4, the debiasing approach naturally adapts to high-dimensional GLMs.

Algorithm 6 The data-splitting method for linear models in the high-dimensional setting.

1. Split the data set into two equal-sized halves $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$.
2. Construct the normalized debiased Lasso estimator on each part of the data.
   (a) Calculate the Lasso estimators $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ via the optimization problem (18).
   (b) Estimate $\widehat{\Theta}^{(1)}$ and $\widehat{\Theta}^{(2)}$ following Algorithm 5. For $j \in [p]$, let
   $$\widehat{\sigma}_j^{2(1)} = \left(\widehat{\Theta}^{(1)}\widehat{\Sigma}^{(1)}\widehat{\Theta}^{(1)\top}\right)_{jj}, \quad \widehat{\sigma}_j^{2(2)} = \left(\widehat{\Theta}^{(2)}\widehat{\Sigma}^{(2)}\widehat{\Theta}^{(2)\top}\right)_{jj}, \tag{24}$$
   in which $\widehat{\Sigma}^{(1)}$ and $\widehat{\Sigma}^{(2)}$ are the sample covariance matrices of $X^{(1)}$ and $X^{(2)}$, respectively.
   (c) Calculate the debiased Lasso estimators $\widehat{\beta}^{(1,d)}$ and $\widehat{\beta}^{(2,d)}$ following Equation (19). For $j \in [p]$, calculate the mirror statistic M_j based on
   $$T_j^{(1)} = \widehat{\beta}_j^{(1,d)}/\widehat{\sigma}_j^{(1)}, \quad T_j^{(2)} = \widehat{\beta}_j^{(2,d)}/\widehat{\sigma}_j^{(2)}. \tag{25}$$
3. Select the features using Algorithm 1.

4.1.2 Theoretical justification of the data-splitting method

We require the following assumptions.

Assumption 4.1.

(1) The sparsity conditions.

(a) The sparsity condition on Θ: $s = \max_{i\in[p]} |\{j \in [p],\ \Theta_{ij} \neq 0\}| = o(\sqrt{n}/\log p)$.

(b) The sparsity condition on the number of signals: $p_1 = |\{j \in [p], \beta_j^\star \neq 0\}| = o(\sqrt{n}/\log p)$.

(2) Requirements for the design matrix X.

(a) The rows of $X\Theta^{1/2}$ are sub-Gaussian.

(b) $1/C \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq C$, for some constant $C > 0$.

(3) The sample size requirement: $\sqrt{n}/\log p \to \infty$.

Remark 4.3. In contrast to the moderate-dimensional setting, we require proper sparsity conditions on the true coefficient vector $\beta^\star$ as well as the precision matrix Θ. The former is required to ensure that the Lasso estimator enjoys a fast convergence rate (Bickel et al., 2009). The latter also appears in Javanmard and Montanari (2013) and Van de Geer et al. (2014), which implies that $||\widehat{\Theta} - \Theta||_\infty = o_p(1/\sqrt{\log p})$. The two sparsity conditions, along with proper conditions on the design matrix, ensure that the bias term Δ vanishes asymptotically in the sense that $||\Delta||_\infty = O_p(p_1 \log p/\sqrt{n}) = o_p(1)$.

The following proposition shows that the data-splitting method achieves an asymptotic FDR control for high-dimensional linear models.

Proposition 4.1. *For any given FDR control level $q \in (0, 1)$, we assume that the pointwise limit $FDP^\infty(t)$ of $FDP(t)$ exists for all $t > 0$, and there is a $t_q > 0$ such that $FDP^\infty(t_q) \leq q$. Then, under Assumption 4.1, we have*

$$\limsup_{n,p\to\infty} \mathbb{E}\left[\frac{\#\{j : j \in S_0, j \in \widehat{S}_{\tau_q}\}}{\#\{j : j \in \widehat{S}_{\tau_q}\} \vee 1}\right] \leq q,$$

using the data-splitting method outlined in Algorithm 6.

4.2 Generalized linear models

4.2.1 Construction of the mirror statistics

In this section, we adapt the debiasing approach to the GLM defined in (10). The Lasso estimator of the true coefficient vector $\beta^\star$ in (10) is

$$\widehat{\beta}(y, X; \lambda) = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{\frac{1}{2n}\sum_{i=1}^n \ell(y_i, x_i^\intercal\beta) + \lambda||\beta||_1\right\}. \tag{26}$$

We refer to $\ell(u, v) = -uv + \rho(v)$ as the loss function associated with the GLM, which is essentially the negative log-likelihood up to an additive constant.

For the ease of presentation, we introduce the following notations. The first and second derivatives of $\ell(u, v)$ with respect to v are denoted as $\dot{\ell}(u, v)$ and $\ddot{\ell}(u, v)$, respectively. The gradient and the Hessian of $\ell(y, x^\intercal\beta)$ with respect to β are denoted as $\dot{\ell}_\beta(y, x)$ and $\ddot{\ell}_\beta(y, x)$, respectively.[2] For

[2]To clarify, both $\dot{\ell}(u, v)$ and $\ddot{\ell}(u, v)$ are scalar, whereas $\dot{\ell}_\beta(y, x)$ is a $p \times 1$ vector and $\ddot{\ell}_\beta(y, x)$ is a $p \times p$ matrix.

a general mapping g defined on an arbitrary data point (y, x), let $P_n g = \sum_{i=1}^n g(y_i, x_i)/n$ and $Pg = \mathbb{E}[P_n g]$, in which the expectation is taken with respect to the randomness in both the response variable y_i and the features x_i. Let W_β be a $n \times n$ diagonal matrix with $W_{i,i}^2 = \ddot{\rho}(x_i^\intercal \beta)$. Then the sample version of the Hessian matrix can be written as $P_n \ddot{\ell}_\beta = X_\beta^\intercal X_\beta / n$, in which $X_\beta = W_\beta X$ is the weighted design matrix.

We consider the following debiased Lasso estimator for high-dimensional GLMs (Van de Geer et al., 2014),

$$\widehat{\beta}^d = \widehat{\beta} - \widehat{\Theta} P_n \dot{\ell}_{\widehat{\beta}}. \tag{27}$$

This is a natural generalization of the debiased Lasso estimator for linear models (see Equation (19)), where $\widehat{\Theta}$ serves as the decorrelating matrix D, and $\dot{\ell}_{\widehat{\beta}}(y, x)$ simplifies to $-x(y - x^\intercal \widehat{\beta})$ in the linear model. Let $\Sigma = \mathbb{E}[X_{\beta^\star}^\intercal X_{\beta^\star}]/n$ be the population Hessian matrix evaluated at the true coefficient vector $\beta^\star$. Similar to the linear model, we set $\widehat{\Theta}$ as an estimator of $\Theta = \Sigma^{-1}$, which is again constructed via regularized node-wise regression (see Algorithm 5). By analogy with the linear model, we have a similar decomposition as Equation (20),

$$\sqrt{n}(\widehat{\beta}_j^d - \beta_j^\star) = Z_j + \Delta_j, \quad \text{for } j \in [p], \tag{28}$$

in which Z_j is the asymptotically dominant term defined as below,

$$Z_j = -\sqrt{n} \Theta_{j,\cdot} P_n \dot{\ell}_{\beta^\star} = -\sqrt{n} \sum_{i=1}^n \Theta_{j,\cdot} x_i [-y_i + \dot{\rho}(x_i^\intercal \beta^\star)]. \tag{29}$$

$\Theta_{j,\cdot}$ denotes the j-th row of Θ. One major difference between GLMs and linear models is that, conditioning on the design matrix, Z_j is not exactly normal but only asymptotically normal by the central limit theorem. Fortunately, we can easily quantify the discrepancy between the law of Z_j and the normal distribution using the Berry-Essen theorem. For the bias term Δ, we show that it can be asymptotically ignored under proper conditions.

The decomposition in Equation (28) suggests a normalized debiased Lasso estimator as below,

$$T_j = \widehat{\beta}_j^d / \widehat{\sigma}_j \quad \text{with} \quad \widehat{\sigma}_j^2 = (\widehat{\Theta} P_n \dot{\ell}_{\widehat{\beta}} \dot{\ell}_{\widehat{\beta}}^\top \widehat{\Theta}^\top)_{jj}, \quad \text{for } j \in [p], \tag{30}$$

in which $\widehat{\sigma}_j^2$ is a consistent estimator of the asymptotic variance

$$\sigma_j^2 = (\Theta \mathbb{E}[P_n \dot{\ell}_{\beta^\star} \dot{\ell}_{\beta^\star}^\top] \Theta)_{jj} = (\Theta \Sigma \Theta)_{jj} = \Theta_{jj}.$$

The symmetric requirement in Condition 2.1 is satisfied since for a null feature $j \in S_0$, T_j is asymptotically centered at 0. A summary of the data-splitting method for high-dimensional GLMs is given in Algorithm 7.

4.2.2 Theoretical justification of the data-splitting method

We require the following assumptions.

Assumption 4.2.

(1) The sparsity conditions.

- *(a) The sparsity condition on Θ.* $s = \max_{i \in [p]} |\{j \in [p],\ \Theta_{ij} \neq 0\}| = o(\sqrt{n}/\log p)$.
- *(b) The sparsity condition on the number of signals.* $p_1 = |\{j \in [p], \beta_j^\star \neq 0\}| = o(\sqrt{n}/\log p)$.

Algorithm 7 The data-splitting method for GLMs in the high-dimensional setting.

1. Split the data set into two equal-sized halves $(y^{(1)}, X^{(1)})$ and $(y^{(2)}, X^{(2)})$.
2. Construct the normalized debiased Lasso estimator on each part of the data.
   (a) Calculate the Lasso estimators $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ via the optimization problem (26).
   (b) Estimate $\widehat{\Theta}^{(1)}$ and $\widehat{\Theta}^{(2)}$ following Algorithm 5. For $j \in [p]$, let
   $$\widehat{\sigma}_j^{2(1)} = \left(\widehat{\Theta}^{(1)} P_n \dot{\ell}_{\widehat{\beta}^{(1)}} \dot{\ell}_{\widehat{\beta}^{(1)}}^\top \widehat{\Theta}^{(1)\top}\right)_{jj}, \quad \widehat{\sigma}_j^{2(2)} = \left(\widehat{\Theta}^{(2)} P_n \dot{\ell}_{\widehat{\beta}^{(2)}} \dot{\ell}_{\widehat{\beta}^{(2)}}^\top \widehat{\Theta}^{(2)\top}\right)_{jj}. \tag{31}$$
   (c) Calculate the debiased Lasso estimators $\widehat{\beta}^{(1,d)}$ and $\widehat{\beta}^{(2,d)}$ following Equation (27). For $j \in [p]$, calculate the mirror statistic M_j based on
   $$T_j^{(1)} = \widehat{\beta}_j^{(1,d)} / \widehat{\sigma}_j^{(1)}, \quad T_j^{(2)} = \widehat{\beta}_j^{(2,d)} / \widehat{\sigma}_j^{(2)}. \tag{32}$$
3. Select the features using Algorithm 1.

(2) Requirements for the design matrix X and the weighted design matrix $X_{\beta^\star}$.

(a) There exists a constant $C_1 > 0$ such that

$$||X||_\infty \le C_1, \quad ||X\beta^\star||_\infty \le C_1, \quad ||X_{\beta^\star}||_\infty \le C_1, \quad ||X_{\beta^\star,-j}\gamma_j||_\infty \le C_1,^{3} \quad \forall j \in [p]. \tag{33}$$

(b) $1/C_2 \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le C_2$, for some constant $C_2 > 0$.

(3) The regularity conditions on the link function ρ.

(a) $\ddot{\rho}(v)$ is Lipschitz continuous for $|v| \le C_1$.

(b) $|\dot{\rho}(v)|$ and $|\ddot{\rho}(v)|$ are upper bounded for $|v| \le C_1$.

(4) Requirements for the sample size: $\sqrt{n}/\log p \to \infty$.

Remark 4.4. For mathematical convenience, we consider bounded (weighted) design matrices in this paper, although the theoretical results can be possibly generalized to more general cases such as sub-Gaussian (weighted) design matrices. Similar regularity conditions on the link function ρ also appears in Van de Geer et al. (2014), and hold for popular GLMs including the logistic regression model, the Poisson regression model, and the negative binomial regression model.

The following propositions show that the bias term Δ vanishes asymptotically, and the data-splitting method achieves an asymptotic FDR control for high-dimensional GLMs.

Proposition 4.2. *Under Assumption 4.2, we have $||\Delta||_\infty = o_p(1)$.*

Proposition 4.3. *For any given FDR control level $q \in (0,1)$, we assume that the pointwise limit $FDP^\infty(t)$ of $FDP(t)$ exists for all $t > 0$, and there is a $t_q > 0$ such that $FDP^\infty(t_q) \le q$. Then, under Assumption 4.2, we have*

$$\limsup_{n,p\to\infty} \mathbb{E}\left[\frac{\#\{j : j \in S_0, j \in \widehat{S}_{\tau_q}\}}{\#\{j : j \in \widehat{S}_{\tau_q}\} \vee 1}\right] \le q,$$

using the data-splitting method outlined in Algorithm 7.

[3] γ_j corresponds to the coefficient vector of the best linear predictor of $X_{\beta^\star,j}$ using $X_{\beta^\star,-j}$. More precisely, we have $\gamma_j = \arg\min_{\gamma\in\mathbb{R}^{p-1}} \mathbb{E}\left[||X_{\beta^\star,j} - X_{\beta^\star,-j}\gamma||_2^2\right]$.

4.3 Comparison with existing methods

The knockoff filter is a class of recently developed methods, which achieves the FDR control by creating knockoff features in a similar spirit as spike-ins in biological experiments. The knockoff filter does not require calculating individual p-values, and can be applied to fairly general settings without having to know the underlying true relationship between the response variable and the associated features. In particular, the model-X knockoff filter (Candes et al., 2018), as well as some further developments including the DeepPINK filter (Lu et al., 2018) and the conditional knockoff filter (Huang and Janson, 2019), can be applied to select features for an arbitrary GLM. Compared to the data-splitting method, one major limitation of the knockoff filter is that it requires the knowledge of the joint distribution of the features. In addition, we empirically observe that the power of the knockoff filter can deteriorate rapidly as correlations among features increase.

BHq is also potentially applicable for the FDR control in GLMs once we obtain the asymptotic p-value for each feature based on the asymptotic normality of the debiased Lasso estimator. Developments along this line include Javanmard and Javadi (2019) and Ma et al. (2020). The former focuses on high-dimensional linear models, while the latter focuses on high-dimensional logistic regression models. Although both the data-splitting method and BHq rely on the asymptotic property of the debiased Lasso estimator, the symmetry requirement is less stringent and more likely to be satisfied in finite-sample cases compared to the normality requirement. In particular, we empirically observe that the scale of the normalized debiased Lasso estimator, i.e., the scale of the Z-score, can be quite off compared to the scale of the standard normal distribution. We note that if the scaling factor is under-estimated, BHq is at the risk of losing the FDR control, since the resulting p-values of the null features will skew to the left. On the other hand, if the scaling factor is over-estimated, BHq can be over conservative, leading to a possible power loss. In contrast, the data-splitting method is scale free, i.e., the scaling factor does not materially change the selection result, thus is potentially more robust compared to BHq.

Figure 1 illustrates the above discussion in a logistic regression model with $n = 250$, $p = 500$ and $p_1 = 10$. Detailed algorithmic settings can be found in the Figure caption. We obtain the normalized debiased Lasso estimator T_j^{BHq} and T_j^{DS} via the method proposed in Ma et al. (2020) and Algorithm 7, respectively. We see that, for BHq, the scale of T_j^{BHq} is much smaller compared to the scale of the standard normal distribution, thus the resulting p-values of the null features are significantly right-skewed. In contrast, for the data-splitting method, the symmetry requirement in Condition 2.1 is well satisfied.

5 Numerical Illustrations

To remind the readers, we use the following abbreviations DS, MDS, BHq, GM, and Knockoff to denote the single data-splitting method, the multiple data-splitting method, the Benjamini-Hochberg procedure, the Gaussian mirror method, and the model-X knockoff filter,[4] respectively. We will clarify the exact implementations of these methods for different numerical examples. For MDS, we repeat the selection for 50 times and aggregate the results using Algorithm 2. In addition, for all the synthetic examples except for the linear model in Section 5.2.1, we keep $|\beta_j^\star|$ the same across relevant features $j \in S_1$, and randomly generate their signs with equal probability. The elements of S_1 are randomly drawn from $\{1, \ldots, p\}$. Furthermore, with a bit abuse of terminology, we refer to $|\beta_j^\star|$ for $j \in S_1$ as the signal strength throughout these synthetic examples. For all the examples in Section 5, we construct the mirror statistic following Equation (4) with $f(u, v) = uv$. The designated FDR control level is set to be $q = 0.1$ henceforth.

[4]We use the R package *knockoff* to implement the model-X knockoff filter. See https://cran.r-project.org/web/packages/knockoff for the documentation.

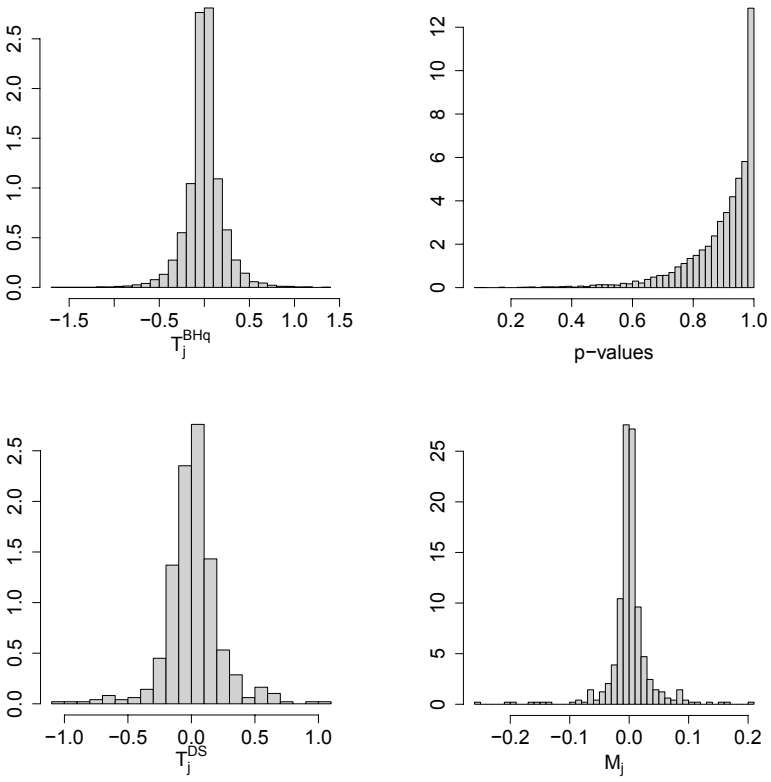

Figure 1: A logistic regression model with $n = 250$, $p = 500$ and $p_1 = 10$. Each row of the design matrix is generated from $N(0, I_p)$. The true coefficient for $j \in S_1$ is set to be $\beta_j^\star = \pm 4$ with equal probability. Top left: The normalized debiased Lasso estimators T_j^{BHq} of the null features. Top right: The asymptotic p-values of the null features. Both histograms are generated based on 20 independent runs of the algorithm in Ma et al. (2020). Bottom left: The normalized debiased Lasso estimators T_j^{DS} of the null features. Bottom right: The mirror statistics of the null features. Both histograms are generated based on a single run of Algorithm 7.

5.1 The moderate-dimensional setting

5.1.1 Logistic regression

We consider two moderate-dimensional settings for the logistic regression model. The first setting is the classical small-n-and-p setting, in which sample size $n = 500$ and dimension $p = 60$, so that the dimension-to-sample ratio of $\kappa = p/n = 0.12$. The second setting concerns with the large-n-and-p setting, in which $n = 3000$, $p = 500$, and $\kappa = 1/6$. The number of relevant features $p_1 = p - p_0$ is 30 in the small-n-and-p setting, and 50 in the large-n-and-p setting. In both settings, we consider six competing methods based on the MLE, including DS, MDS, GM, BHq along with its adjusted version ABHq, and Knockoff. The implementation details of DS and GM are given in Algorithm 3 and 4, respectively. BHq utilizes the classical asymptotic p-values calculated via the Fisher information, whereas ABHq is based on the adjusted asymptotic p-values derived recently by Sur and Candès (2019).

Each row of the design matrix is independently drawn from the multivariate normal distribution $N(0, \Sigma)$ with a Toeplitz correlation structure, i.e., $\Sigma_{ij} = r^{|i-j|}$. The variance of each feature is then standardized to be $1/n$. In Section B.1.1 of the Appendix, we report additional results for different types of covariance matrix Σ, including the case where features have constant pairwise correlation.

In this example, we vary (a) the correlation r; (b) the signal strength $|\beta_j^\star|$ for $j \in S_1$. In the small-n-and-p setting, for scenario (a), we fix the signal strength at $|\beta_j^\star| = 6.5$ for $j \in S_1$, and vary the correlation r from 0.0 to 0.4, whereas for scenario (b), we fix the correlation at $r = 0.2$, and vary the signal strength from 4.5 to 6.5. In the large-n-and-p setting, for scenario (a), we fix the signal strength at $|\beta_j^\star| = 11$ for $j \in S_1$, and vary the correlation r from 0.0 to 0.4, whereas for scenario (b), we fix the correlation at $r = 0.2$, and vary the signal strength from 8 to 12.

The empirical FDRs and powers of different methods in the small-n-and-p setting are summarized in Figure 2. The FDRs of the six competing methods are under control across all settings. In terms of the power, BHq is the leading method, and performs the best in all cases. MDS is the second best method, having a slightly lower power than BHq. We observe that ABHq is less powerful than BHq, indicating that the asymptotics for the p-value adjustment is not ready to kick in when the sample size n and the dimension p are relatively small.

After adding p knockoff variables created by the second-order method with a James-Stein-type shrinkage applied to the estimated covariance matrix (Barber and Candès, 2015), the sample covariance matrix of the augmented set of features appears almost singular; thus the resulting MLEs are unstable. To overcome this issue, we multiply the vector s output from the SDP program (see Equation (2.4) in Barber and Candès (2015)) by a factor 0.9. The power of the resulting knockoff filter, as shown in Figures 2 and 3, is still unsatisfactory. We also tested out the recently proposed conditional knockoff filter (Huang and Janson, 2019), which is free of the covariance matrix estimation, but observed no improvements in power.

Except for MDS, all the perturbation-based methods, such as GM, DS, and Knockoff are not as powerful as the p-value-based methods. A possible reason is that when p and n are small, the types of perturbations in these methods may have diluted the signal too much. More interestingly, however, MDS gains back almost all the lost power due to perturbations without sacrificing FDR controls.

The empirical FDRs and powers of different methods in the large-n-and-p setting are summarized in Figure 3. We see that BHq loses the FDR control because the classical asymptotic p-value of the null feature, calculated based on the Fisher information, is non-uniform and skew to the left. In contrast, ABHq still controls the FDR well and enjoys the best power, which verifies the adjusted asymptotic distribution of the MLE derived in Sur and Candès (2019). MDS and GM also perform competitively, and have similar but slightly lower power compared to ABHq. GM shows much improved performances compared with the small n-and-p setting. Knockoff performs poorly for the same reason as mentioned above.

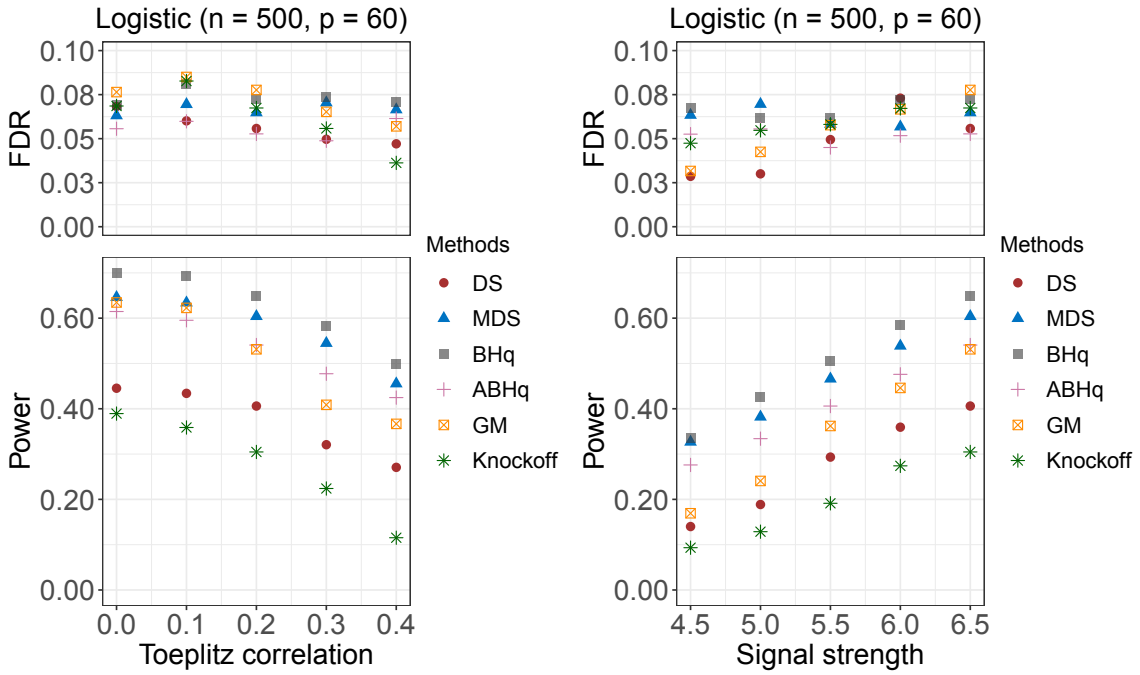

Figure 2: Empirical FDRs and powers for the logistic regression model in the small-n-and-p setting. Each row of the design matrix is first independently drawn from $N(0, \Sigma)$ with a Toeplitz correlation structure, i.e., $\Sigma_{ij} = r^{|i-j|}$. The variance of each feature is then standardized to be $1/n$. The power is defined as the ratio of the identified versus all relevant features. In the left panel, we fix the signal strength at $|\beta_j^\star| = 6.5$ for $j \in S_1$ and vary the correlation r. In the right panel, we fix the correlation at $r = 0.2$ and vary the signal strength. The number of relevant features is 30 across all settings, and the designated FDR control level is $q = 0.1$. Each dot represents the average from 50 independent runs.

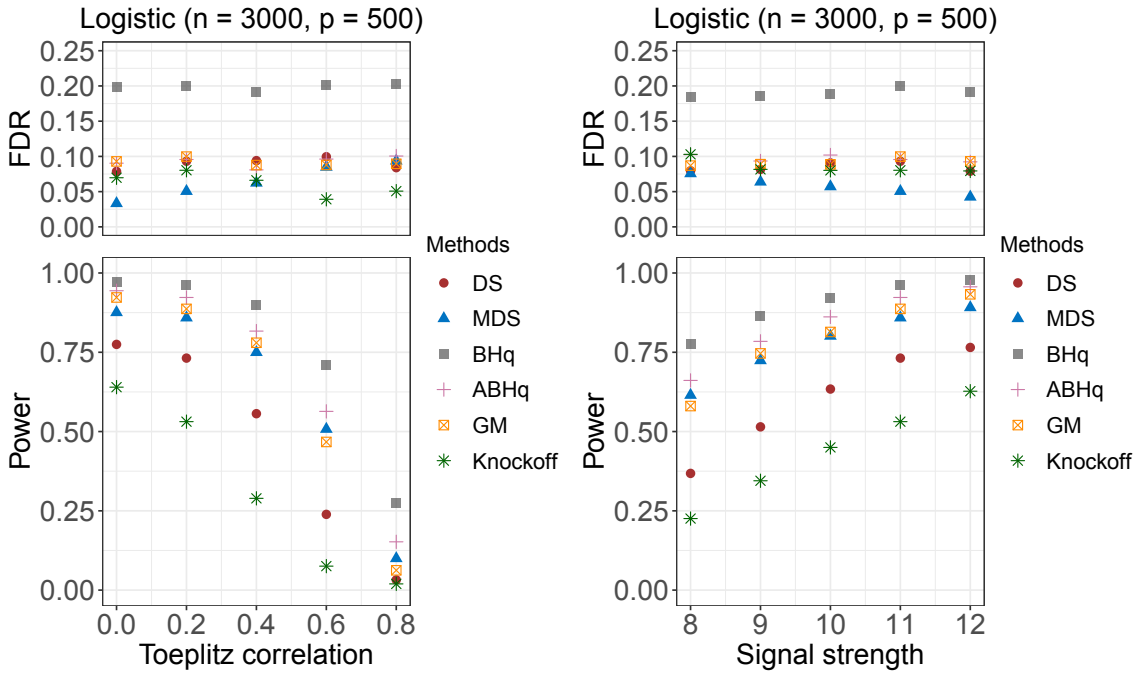

Figure 3: Empirical FDRs and powers for the logistic regression model in the large-n-and-p setting. The design matrix is simulated similarly as per Figure 2. In the left panel, we fix the signal strength at $|\beta_j^\star| = 11$ for $j \in S_1$, and vary the correlation r. In the right panel, we fix the correlation at $r = 0.2$, and vary the signal strength. The number of relevant features is 50 across all settings, and the designated FDR control level is $q = 0.1$. Each dot represents the average from 50 independent runs.

5.1.2 Negative binomial regression

We consider a negative binomial regression model, in which the dispersion parameter is 2, i.e., the target number of successful trials is 2. We set sample size $n = 3000$, and dimension $p = 500$, resulting in a dimension-to-sample ratio of $\kappa = 1/6$. We simulate the design matrix the same way as in Section 5.1.1, and vary (a) the correlation r; (b) the signal strength $|\beta_j^\star|$ for $j \in S_1$. In scenario (a), we fix the signal strength at $|\beta_j^\star| = 6$ for $j \in S_1$, and vary the correlation r from 0.1 to 0.5. In scenario (b), we fix the correlation at $r = 0.2$, and vary the signal strength from 4 to 8. The number of relevant features is 50 across all settings, i.e., $p_1 = 50$ and $p_0 = 450$. In Section B.1.2 of the Appendix, we report additional results for the case where features have constant pairwise correlation.

We consider five competing methods based on the MLE, including DS, MDS, BHq, GM and Knockoff. The implementation details of DS and GM are given in Algorithms 3 and 4, respectively. BHq is based upon the classical asymptotic p-values calculated via the Fisher information. Although we expect such p-values to be non-uniform for the null features, the exact asymptotic distribution of the MLE under this moderate-dimensional setting has not been derived in the literature, thus no proper adjustment of the p-values exists to the best of our knowledge. Knockoff is implemented in the same way as in Section 5.1.1 to overcome the degeneracy issue.

The empirical FDRs and powers of different methods are summarized in Figure 4. We see that BHq is the only method losing the FDR control, because of the non-uniformity (skew to the left) of the p-values for the null features. Among the methods with FDR control, GM and MDS consistently perform the best over different levels of correlation and signal strength. MDS has a silghtly lower power but also a lower FDR compared with GM, and is significantly better than DS in the sense that it simultaneously reduces the FDR and boosts the power. Knockoff has the lowest power among all competing methods for the same reason as discussed in Section 5.1.1.

5.2 The high-dimensional setting

5.2.1 Linear regression

We consider the Gaussian linear model $y = X\beta^\star + \epsilon,\ \epsilon \sim N(0, I_n)$, with sample size $n = 800$ and dimension $p = 2000$. Each row of the design matrix is drawn independently from the multivariate Gaussian $N(0, \Sigma)$ with a Toeplitz correlation structure. More precisely, Σ is a blockwise diagonal matrix, consisting of 10 identical unit diagonal Toeplitz matrices. The detailed formula involves a correlation factor $r \in (0, 1)$, and is given in Section B.2.1 of the Appendix. Features are more correlated with a larger r. We independently sample $\beta_j^\star$ for $j \in S_1$ from a centered normal distribution, of which the standard deviation is referred to as the signal strength.

In this example, we vary (a) the correlation factor r; (b) the signal strength. In scenario (a), we fix the signal strength at $6\sqrt{\log p/n}$, and vary the correlation factor r from 0.0 to 0.8, whereas in scenario (b), we fix the correlation factor at $r = 0.6$, and vary the signal strength from 2 to 10 up to a multiplicative constant $\sqrt{\log p/n}$. The number of relevant features is 70 across all settings, i.e., $p_1 = 70$. In Section B.2.1 of the Appendix, we report additional results for the case where features have constant pairwise correlation.

We consider four competing methods, including DS, MDS, the BHq procedure outlined in Javanmard and Javadi (2019), and Knockoff. DS and MDS are based on the debiased Lasso estimator, with implementation details given in Algorithm 6. BHq utilizes the same debiasing approach, and estimates the noise level using scaled Lasso. For Knockoff, we use the second-order method to create multivariate normal knockoffs.

The empirical FDRs and powers of different methods are summarized in Figure 5. The FDRs of the four competing methods are under control across all settings. In particular, MDS has much lower FDRs than DS, also significantly lower than the nominal level. In terms of the power, Knockoff

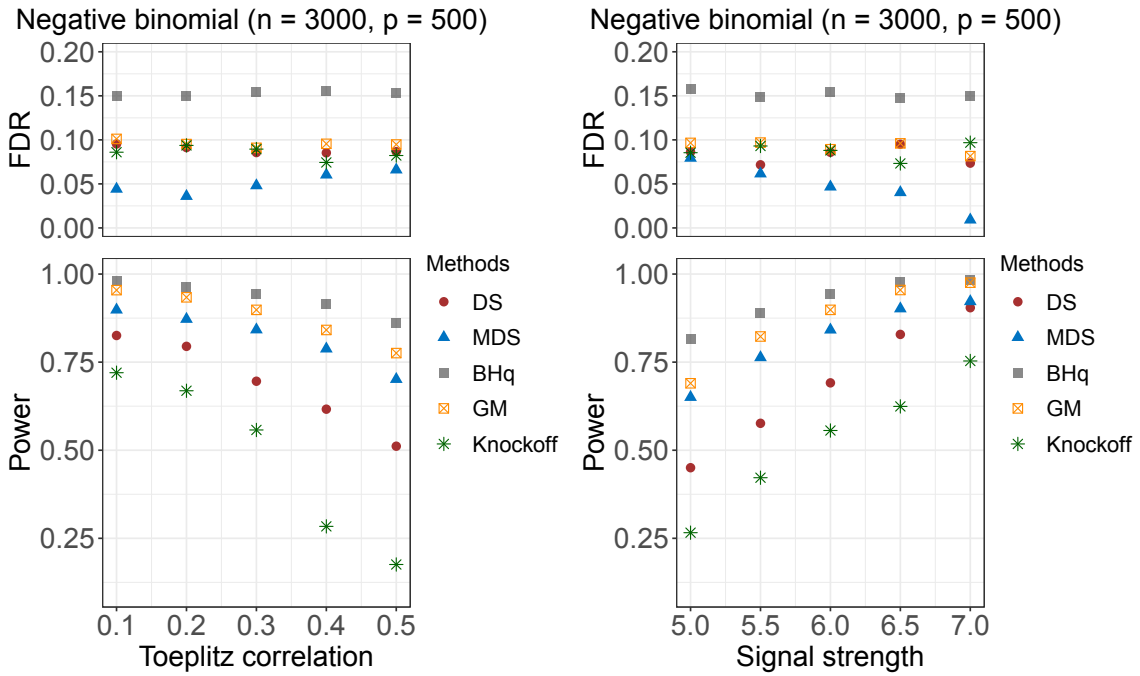

Figure 4: Empirical FDRs and powers for the negative binomial regression model. The design matrix is simulated similarly as per Figure 2. In the left panel, we fix the signal strength at $|\beta_j^\star| = 6$ for $j \in S_1$, and vary the correlation r. In the right panel, we fix the correlation at $r = 0.3$, and vary the signal strength. The number of relevant features is 50 across all setting, and the designated FDR control level is $q = 0.1$, Each dot represents the average from 50 independent runs.

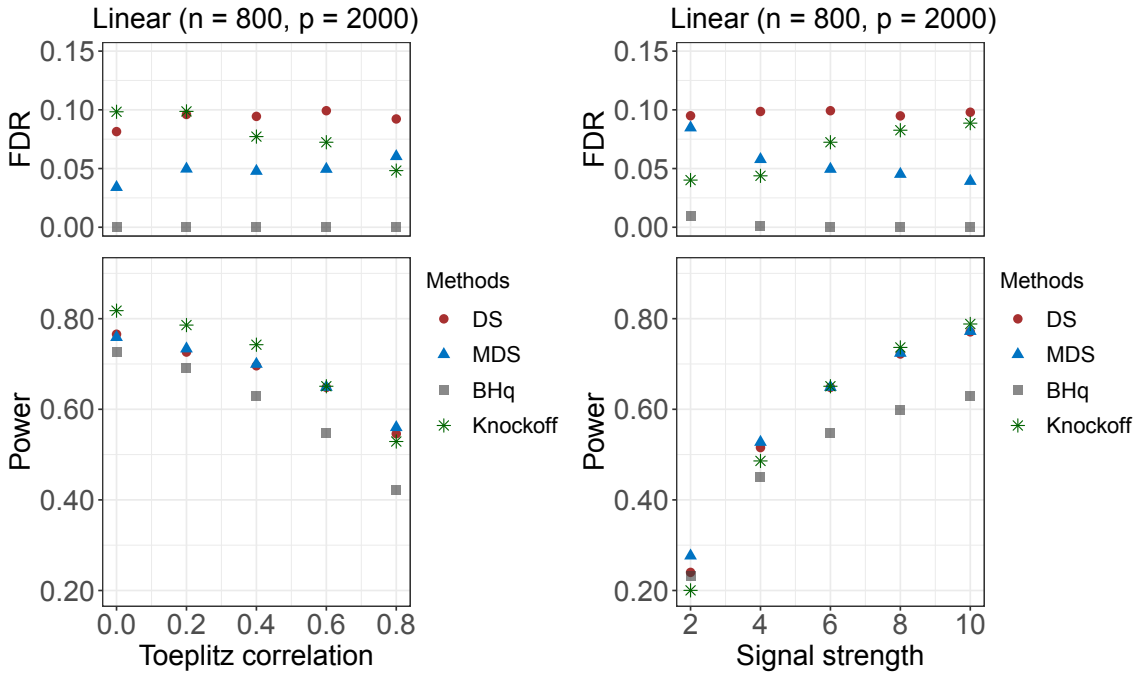

Figure 5: Empirical FDRs and powers for the linear model. Each row of the design matrix is independently drawn from $N(0, \Sigma)$ with Σ being a blockwise diagonal matrix that consists of 10 identical unit diagonal Toeplitz matrices (see Section B.2.1 of the Appendix). $\beta_j^\star$ for $j \in S_1$ are i.i.d. samples from $N(0, s^2)$, where s is referred to as the signal strength. The signal strength along the x-axis in the right panel shows multiples of $\sqrt{\log p/n}$. In the left panel, we fix the signal strength at $6\sqrt{\log p/n}$ and vary the correlation factor r. In the right panel, we fix the correlation factor at $r = 0.6$ and vary the signal strength. The number of relevant features is 70 across all settings, and the designated FDR control level is $q = 0.1$. Each dot in the figure represents the average from 50 independent runs.

achieves the highest power in most cases except when the signal strength is low. We also observe that when the correlation among relevant features increases, such as when the design matrix has constant pairwise correlation (Appendix, Section B.2.1), the power of Knockoff decreases rapidly and can be much lower compared with the other three methods. DS and MDS also perform competitively and consistently enjoy a higher power than BHq over different levels of correlation and signal strength. We see that the empirical FDRs of BHq are nearly zero across all settings. One possible reason is that the scaled Lasso over estimated the noise level, thus the p-values largely skew to the right and make the procedure too conservative. The numerical results suggest that bypassing the task of estimating the noise level enables the data-splitting methods to gain a substantial advantage over BHq.

5.2.2 Logistic regression

We consider a case with sample size $n = 800$ and dimension $p = 2000$. Each row of the design matrix is drawn independently from $N(\mathbf{0}, \Sigma)$. We consider a similar setup as in Ma et al. (2020), in which $\Sigma = 0.1 \times \Sigma_B$, where Σ_B is the blockwise diagonal matrix introduced in Section 5.2.1 with $r = 0.1$. In this example, we vary (a) the sparsity level p_1, i.e., the number of relevant features; (b) the signal strength $|\beta_j^\star|$ for $j \in S_1$. In scenario (a), we fix the signal strength at $|\beta_j^\star| = 4$ for $j \in S_1$, and vary the sparsity level p_1 from 40 to 80. In scenario (b), we fix the sparsity level at $p_1 = 60$,

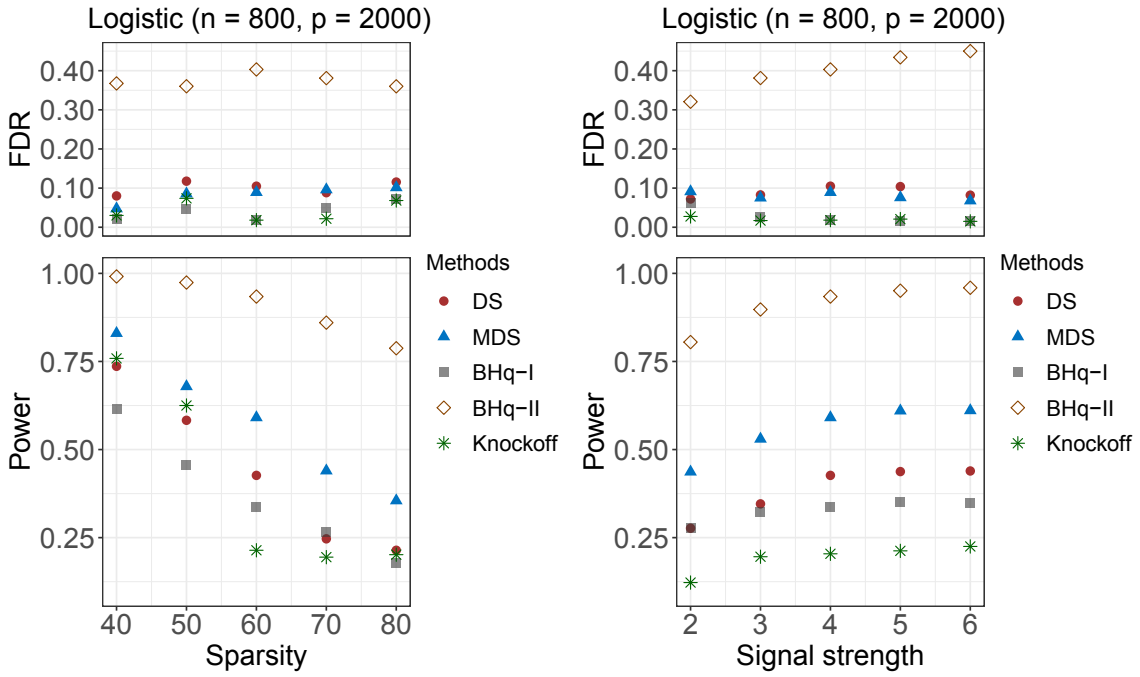

Figure 6: Empirical FDRs and powers for the logistic regression model. The rows of the design matrix are i.i.d. samples from $N(0_p, 0.1 \times \Sigma_B)$, where Σ_B is a blockwise diagonal matrix that consists of 10 identical unit diagonal Toeplitz matrices (see Section B.2.1 of the Appendix with $r = 0.1$). In the left panel, we fix the signal strength at $|\beta_j^\star| = 4$ for $j \in S_1$, and vary the sparsity level p_1. In the right panel, we fix the sparsity level at $p_1 = 60$, and vary the signal strength. The designated FDR level is $q = 0.1$. Each dot represents the average of 50 independent runs.

and vary the signal strength from 2 to 6.

We consider five competing methods, including DS, MDS, two BHq procedures (BHq-I and BHq-II), and Knockoff. DS and MDS are based on the debiased Lasso estimator, with implementation details given in Algorithm 7. BHq-II uses the same debiasing approach, whereas BHq-I, which corresponds to the method proposed in Ma et al. (2020), utilizes a different debiasing approach. For Knockoff, we use the second-order method to create multivariate normal knockoffs.

The empirical FDRs and powers of different methods are summarized in Figure 6. We see that all methods except BHq-II control the FDR successfully. In terms of power, MDS is the leading method across different levels of sparsity and signal strength, and enjoys a significant improvement over DS. Even DS appears to be more powerful than BHq-I, suggesting that the p-values constructed following Ma et al. (2020) can be highly non-informative (skew to the right) in finite-sample cases. Knockoff performs competitively when the number of relevant features is small (e.g., $p_1 \leq 50$), but can potentially suffer when the relevant features become denser, i.e., p_1 gets larger.

5.3 Real data application

Compared with traditional bulk RNA sequencing technologies, single-cell RNA sequencing (scRNAseq) allows researchers to examine the sequence information of each individual cell, which promises to lead to new biological discoveries ranging from cancer genomics to metagenomics. In this section, we consider the task of selecting relevant genes with respect to the glucocorticoid response in a human breast cancer cell line, using the scRNAseq data in Hoffman et al. (2020). A total of 400 T47D A12 human breast cancer cells were treated with 100 nM synthetic glucocorticoid

dexamethasone (Dex) at 1, 2, 4, 8, and 18h time points. An scRNASeq experiment was performed at each time point, which results in a total of 2,000 samples of gene expression for the treatment group. For the control group, there are 400 vehicle-treated control cells. An scRNAseq experiment was performed at the 18h timepoint to obtain the corresponding profile of gene expression. After proper normalization, the final scRNAseq dataset[5] contains 2,400 samples, each with 32,049 gene expressions. To further reduce the dimensionality, we first screen out the genes detected in fewer than 10% of cells, and then pick up the top 500 most variable genes following Hoffman et al. (2020).

We consider a logistic regression model with $n = 2,400$ and $p = 500$. Since the MLE does not exist for this dataset, we cannot use the method in Sur and Candès (2019) to obtain relevant p-values. Instead, we use the debiased Lasso estimator and apply DS outlined in Algorithm 7 to this dataset. As the sample size n is larger than the dimension p, we directly estimate the precision matrix Θ (see Equation (27) and the discussion therein) by inverting the sample Hessian matrix $\widehat{\Sigma}$, instead of using the regularized node-wise regression outlined in Algorithm 5. Two other methods are tested out as well, including the BHq procedure outlined in Ma et al. (2020) and Knockoff.

With the nominal FDR control level set at $q = 0.1$, MDS performs significantly more powerful than the other two competing methods. More precisely, MDS stably selects approximately 30 genes (see Table 1), in the sense that the selection results obtained via two independent runs of MDS only differ by one or two genes. In contrast, Knockoff approximately selects 13 genes, forming a subset of the genes selected by MDS (see Table 1). BHq only selects 1 gene, RPL10, which is also consistently selected by MDS and Knockoff. Figure 7 demonstrates the sharp difference in the gene expression distributions between the treatment group and the control group for 4 genes: FKBP5, NFKBIA, HSPA1A, and RBM24 selected by MDS, of which the first two are also selected by Knockoff.

The existing literature confirms the interactions between the glucocorticoid receptor (GR) and many genes selected by both MDS and Knokoff, as well as the majority selected only by MDS, thus backing up our selection results to some extent. We highlight the following supporting evidences, and provide a summary of the references associated with the selected genes in Table 1.

- Genes selected by both MDS and Knockoff.

  (i) The SERPINA6 gene is a coding gene for the protein corticosteroid-binding globulin (CBG), which is a major transport protein for glucocorticoids and progestins in the blood of almost all vertebrate species (Zhou et al., 2008). Among its related pathways are Glucocorticoid Pathway (Peripheral Tissue), Pharmacodynamics.

  (ii) The FKBP5 gene encodes the FK506 binding protein 51, a co-chaperone in the heat shock protein 90 (Hsp90) and steroid complex, which regulates GR sensitivity (Nair et al., 1997).

  (iii) The NFKBIA gene is a coding gene for the protein NF-kappa-B inhibitor alpha, and the glucocorticoids are potent inhibitors of nuclear factor kappa B (NF-kappa B) activation (Auphan et al., 1995; Deroo and Archer, 2001).

  (iv) Williamson et al. (2020) showed that activation of PlexinB1 by SEMA3C and SEMA4D promotes nuclear translocation of the GR.

  (v) The HSPB1 gene encodes the heat shock protein 27 (Hsp27), and the synthesis of Hsp27 is regulated by glucocorticoids. The HSPA1A gene also interacts with Hsp70. *In vitro* it acts as an ATP-independent chaperone by inhibiting protein aggregation and by stabilizing partially denatured proteins, which ensures refolding by the Hsp70-complex (Barr and Dokas, 1999).

- Genes selected by MDS only.

[5]The dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE141834.

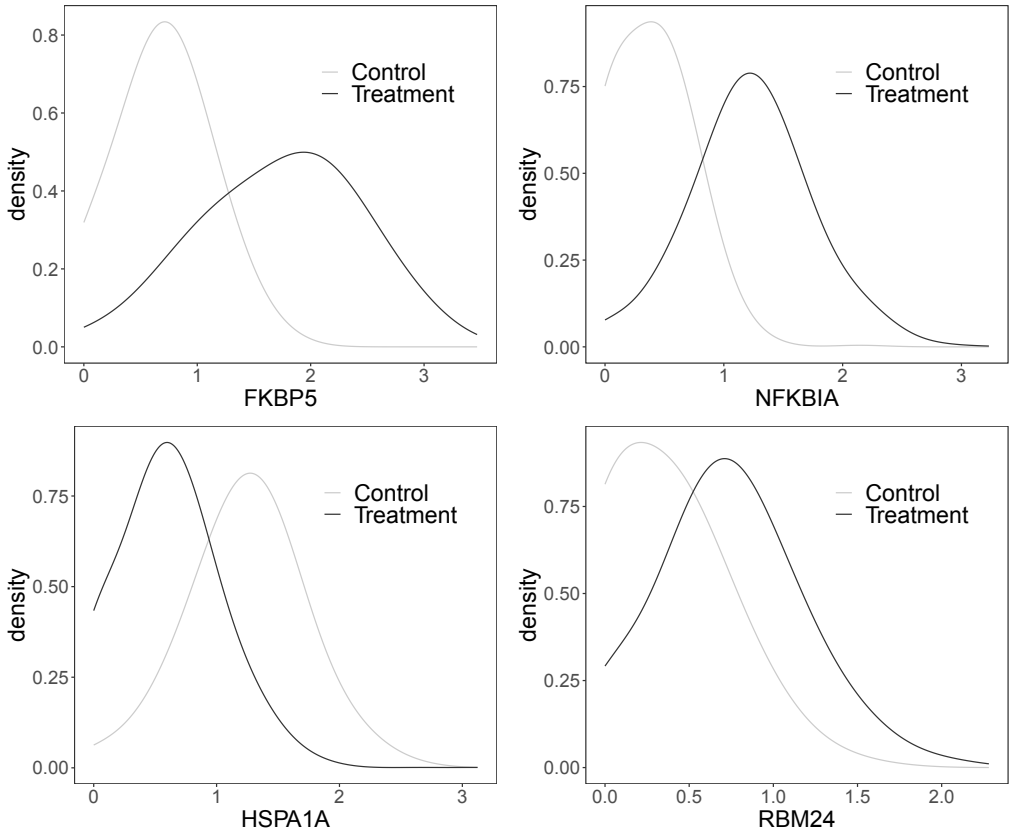

Figure 7: Comparison of the gene expression distributions between the synthetic glucocorticoid dexamethasone treatment group and the vehicle-treated control group for 4 genes. FKBP5 and NFKBIA are selected by both MDS and the model-X knockoff filter, whereas HSPA1A and RBM24 are only selected by MDS.

(i) The HSPA8 gene is a coding gene for the protein Hsc70. The cooperation of Hsc70 with Hsp90 regulates the GR activation and signaling pathway (Furay et al., 2006).

(ii) The HSPA1A gene encodes a 70kDa heat shock protein, which is a member of the heat shock protein 70 (Hsp70) family. Hsp70 promotes GR ligand release and inactivation by inducing partial unfolding (Kirschke et al., 2014).

(iii) Goodman et al. (2016) reported that glucocorticoids, along with mineralocorticoids, effectively expand the cytokeratin-5-positive cell population induced by 3-ketosteroids, which requires induction of the transcriptional repressor BCL6 based on suppression of BCL6.

(iv) The ATF4 gene is a coding gene for the activating transcription factor 4, which is repressed by glucocorticoids and induced by insulin (Adams, 2007).

(v) The IGFBP4 gene encodes the insulin-like growth factor-binding protein 4, regulated by glucocorticoids (e.g. dexamethasone inhibits IGFBP4 expression) (Cheung et al., 1994; Okazaki et al., 1994; Conover et al., 1995).

(vi) The YWHAQ gene encodes the 14–3-3 protein theta, and 14–3-3 protein interacts with the ligand-activated GR (e.g., overexpression of the 143-3 protein enhances the transcriptional activity of glucocorticoid receptor in transient transfection experiments) (Wakui et al., 1997; Zilliacus et al., 2001).

(vii) The DDIT4 gene encodes the DNA damage-inducible transcript 4 protein, and among its inducers are glucocorticoids (Wang et al., 2003; Boldizsár et al., 2006; Wolff et al., 2014).

(viii) The RBM24 gene encodes the RNA-binding protein 24, and is antagonistically regulated by glucocorticoid dexamethasone (Whirledge et al., 2013).

6 Conclusion

We have described a general framework for the task of feature selection in GLMs with FDR asymptotically under control. In particular, we detail the construction of the mirror statistics under two asymptotic regimes, including the moderate-dimensional setting ($p/n \to \kappa \in (0, 1)$) and the high-dimensional setting ($p \gg n$). Compared to BHq, the proposed methodology enjoys a wider applicability and improved robustness thanks to its scale-free property. Compared to the knockoff filter, the proposed methodology does not require the knowledge of the joint distribution of the features, which is the norm for many practical problems, and is less affected by the correlations among the features.

We conclude by pointing out several directions for future work. First, it is of immediate interest to generalize the proposed methods in order to cover the case where the set of explanatory features exhibit a group structure. Second, we would like to investigate the potential applicability of our FDR control framework to dependent observations (e.g., stationary time series data). These two types of data structures appear a lot in practice including genetic studies and financial engineering. Third, moving beyond the parametric models, we would like to consider the FDR control problem in semiparametric single-index models, in which the link function becomes unknown.

Table 1: The references in support of the genes selected by MDS.

Gene	Knockoff	Reference
SERPINA6	✓	GeneCards-SERPINA6, Zhou et al. (2008)
FKBP5	✓	AGCOH-FKBP5, Nair et al. (1997)
NFKBIA	✓	Auphan et al. (1995); Deroo and Archer (2001)
RPL10	✓	Zorzatto et al. (2015)
SEMA3C	✓	Williamson et al. (2020)
HSPB1	✓	Barr and Dokas (1999); Tuckermann et al. (1999)
RBBP7	✓	Jangani et al. (2014)
EIF4EBP1	✓	Watson et al. (2012)
S100A11	✓	String-S100A11, Reeves et al. (2009)
NUPR1	✓	Wikigenes-NUPR1, Mukaida et al. (1994)
MSX2	✓	Jaskoll et al. (1998)
LY6E	✓	–
BLOC1S1	✓	–
HSPA8	✗	Furay et al. (2006)
HSPA1A	✗	Kirschke et al. (2014)
EEF1A1	✗	NCBI-EEF1A1
BCL6	✗	Goodman et al. (2016)
ATF4	✗	Adams (2007)
IGFBP4	✗	Cheung et al. (1994); Okazaki et al. (1994); Conover et al. (1995)
YWHAQ	✗	Wakui et al. (1997); Zilliacus et al. (2001)
DDIT4	✗	Wang et al. (2003); Boldizsár et al. (2006); Wolff et al. (2014)
IRX2	✗	Lambert et al. (2013)
GATA3-AS1	✗	Liberman et al. (2009)
RBM24	✗	Whirledge et al. (2013)
TACSTD2	✗	McDougall (2017)
DSCAM-AS1	✗	Zhao et al. (2016); Chen and Cai (2020)
C1QBP	✗	Sheppard (1994); Cote-Vélez et al. (2008); Zhang et al. (2013)
SNHG19	✗	–
PRR15L	✗	–
RPLP0P6	✗	–
UHMK1	✗	–

The second column indicates whether the gene is also selected by the model-X knockoff filter. There are 6 genes that we do not find direct supporting evidence in the existing literature for their interaction with the glucocorticoid receptor (GR), which might be of interest for further investigations. The red hyperlinks point to the documented information of the corresponding genes in some widely referred databases including GeneCards, Strings, Wikigenes, the National Center for Biotechnology Information (NCBI), and Atlas of Genetics and Cytogenetics in Oncology and Harmatology (AGCOH).

References

Adams, C. M. (2007). Role of the transcription factor ATF4 in the anabolic actions of insulin and the anti-anabolic actions of glucocorticoids. *Journal of Biological Chemistry*, 282(23):16744–16753.

Auphan, N., DiDonato, J. A., Rosette, C., Helmberg, A., and Karin, M. (1995). Immunosuppression by glucocorticoids: inhibition of NF-kappa B activity through induction of I kappa B synthesis. *Science*, 270(5234):286–290.

Azriel, D. and Schwartzman, A. (2015). The empirical distribution of a large number of correlated normal variables. *Journal of the American Statistical Association*, 110(511):1217–1228.

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

Barr, C. S. and Dokas, L. A. (1999). Glucocorticoids regulate the synthesis of HSP27 in rat brain slices. *Brain Research*, 847(1):9–17.

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

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Boldizsár, F., Pálinkás, L., Czömpöly, T., Bartis, D., Németh, P., and Berki, T. (2006). Low glucocorticoid receptor (GR), high Dig2 and low Bcl-2 expression in double positive thymocytes of BALB/c mice indicates their endogenous glucocorticoid hormone exposure. *Immunobiology*, 211(10):785–796.

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

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

Candès, E. J. and Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42.

Chen, H. and Cai, K. (2020). DSCAM-AS1 mediates pro-hypertrophy role of GRK2 in cardiac hypertrophy aggravation via absorbing miR-188-5p. *In vitro cellular and developmental biology. Animal*.

Cheung, P. T., Wu, J., Banach, W., and Chernausek, S. D. (1994). Glucocorticoid regulation of an insulin-like growth factor-binding protein-4 protease produced by a rat neuronal cell line. *Endocrinology*, 135(4):1328–1335.

Conover, C. A., Clarkson, J. T., and Bale, L. K. (1995). Effect of glucocorticoid on insulin-like growth factor (IGF) regulation of IGF-binding protein expression in fibroblasts. *Endocrinology*, 136(4):1403–1410.

Cote-Vélez, A., Pérez-Martínez, L., Charli, J., and Joseph-Bravo, P. (2008). The PKC and ERK/MAPK pathways regulate glucocorticoid action on TRH transcription. *Neurochemical Research*, 33(8):1582.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2020). False discovery rate control via data splitting. *arXiv:2002.08542*.

Deroo, B. J. and Archer, T. K. (2001). Glucocorticoid receptor activation of the I kappa B alpha promoter within chromatin. *Molecular Biology of the Cell*, 12(11):3365–3374.

Furay, A. R., Murphy, E. K., Mattson, M. P., Guo, Z., and Herman, J. P. (2006). Region-specific regulation of glucocorticoid receptor/HSP90 expression and interaction in brain. *Journal of Neurochemistry*, 98(4):1176–1184.

Goodman, C. R., Sato, T., Peck, A. R., Girondo, M. A., Yang, N., Liu, C., Yanac, A. F., Kovatich, A. J., Hooke, J. A., and Shriver, C. D. (2016). Steroid induction of therapy-resistant cytokeratin-5-positive cells in estrogen receptor-positive breast cancer through a BCL6-dependent mechanism. *Oncogene*, 35(11):1373–1385.

Hoffman, J. A., Papas, B. N., Trotter, K. W., and Archer, T. K. (2020). Single-cell RNA sequencing reveals a heterogeneous response to glucocorticoids in breast cancer cells. *Communications Biology*, 3(1):1–11.

Huang, D. and Janson, L. (2019). Relaxing the assumptions of knockoffs by conditioning. *arXiv preprint arXiv:1903.02806*.

Jangani, M., Poolman, T. M., Matthews, L., Yang, N., Farrow, S. N., Berry, A., Hanley, N., Williamson, A. J. K., Whetton, A. D., and Donn, R. (2014). The methyltransferase WBSCR22/Merm1 enhances glucocorticoid receptor function and is regulated in lung inflammation and cancer. *Journal of Biological Chemistry*, 289(13):8931–8946.

Jaskoll, T., Luo, W., and Snead, M. L. (1998). Msx-2 expression and glucocorticoid-induced overexpression in embryonic mouse submandibular glands. *Journal of Craniofacial Genetics and Developmental biology*, 18(2):79–87.

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

Javanmard, A. and Montanari, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1427–1434. IEEE.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Kirschke, E., Goswami, D., Southworth, D., Griffin, P. R., and Agard, D. A. (2014). Glucocorticoid receptor function regulated by coordinated action of the Hsp90 and Hsp70 chaperone cycles. *Cell*, 157(7):1685–1697.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). Bivariate and trivariate normal distributions. *Continuous multivariate distributions*, 1:251–348.

Lambert, W. M., Xu, C., Neubert, T. A., Chao, M. V., Garabedian, M. J., and Jeanneteau, F. D. (2013). Brain-derived neurotrophic factor signaling rewrites the glucocorticoid transcriptome via glucocorticoid receptor phosphorylation. *Molecular and Cellular Biology*, 33(18):3700–3714.

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

Liberman, A. C., Druker, J., Refojo, D., Holsboer, F., and Arzt, E. (2009). Glucocorticoids inhibit GATA-3 phosphorylation and activity in T cells. *The FASEB Journal*, 23(5):1558–1571.

Lu, Y., Fan, Y., Lv, J., and Noble, W. S. (2018). DeepPINK: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 8676–8686.

Ma, R., Cai, T. T., and Li, H. (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *Journal of the American Statistical Association*, pages 1–15.

McDougall, A. R. A. (2017). The role of Trop2 in fetal lung development. *Thesis*.

Mukaida, N., M., M., Ishikawa, Y., Rice, N., Okamoto, S., Kasahara, T., and Matsushima, K. (1994). Novel mechanism of glucocorticoid-mediated gene repression. Nuclear factor-kappa B is target for glucocorticoid-mediated interleukin 8 gene repression. *Journal of Biological Chemistry*, 269(18):13289–13295.

Nair, S. C., Rimerman, R. A., Toran, E. J., Chen, S., Prapapanich, V., Butts, R. N., and Smith, D. F. (1997). Molecular cloning of human FKBP51 and comparisons of immunophilin interactions with HSP90 and progesterone receptor. *Molecular and Cellular Biology*, 17(2):594–603.

Okazaki, R., Riggs, B. L., and Conover, C. A. (1994). Glucocorticoid regulation of insulin-like growth factor-binding protein expression in normal human osteoblast-like cells. *Endocrinology*, 134(1):126–132.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259.

Reeves, E. K. M., Gordish-Dressman, H., Hoffman, E. P., and Hathout1, Y. (2009). Proteomic profiling of glucocorticoid-exposed myogenic cells: time series assessment of protein translocation and transcription of inactive mrnas. *Proteome Science*, 7.

Salehi, F., Abbasi, E., and Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, pages 11982–11992.

Sheppard, K. E. (1994). Calcium and protein kinase C regulation of the glucocorticoid receptor in mouse corticotrope tumor cells. *The Journal of Steroid Biochemistry and Molecular Biology*, 48(4):337–345.

Sun, T. and Zhang, C. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. *Probability Theory and Related Fields*, 175:487558.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized M-estimators in high-dimensions. *IEEE Transactions on Information Theory*, 64(8):5592 – 5628.

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

Tuckermann, J. P., Reichardt, H. M., Arribas, R., Richter, K. H., Schütz, G., and Angel, P. (1999). The DNA binding-independent function of the glucocorticoid receptor mediates repression of AP-1–dependent genes in skin. *The Journal of Cell Biology*, 147(7):1365–1370.

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

Vershynin, R. (2010). *Introduction to the non-asymptotic analysis of random matrices*.

Wakui, H., Wright, A. P. H., Gustafsson, J., and Zilliacus, J. (1997). Interaction of the ligand-activated glucocorticoid receptor with the 14-3-3 eta protein. *Journal of Biological Chemistry*, 272(13):8153–8156.

Wang, Z., Malone, M. H., Thomenius, M. J., Zhong, F., Xu, F., and Distelhorst, C. W. (2003). Dexamethasone-induced gene 2 (Dig2) is a novel pro-survival stress gene induced rapidly by diverse apoptotic signals. *Journal of Biological Chemistry*, 278(29):27053–27058.

Watson, M. L., Baehr, L. M., Reichardt, H. M., Tuckermann, J. P., Bodine, S. C., and Furlow, J. D. (2012). A cell-autonomous role for the glucocorticoid receptor in skeletal muscle atrophy induced by systemic glucocorticoid exposure. *American Journal of Physiology-Endocrinology and Metabolism*, 302(10):E1210–E1220.

Whirledge, S., Xu, X., and Cidlowski, J. A. (2013). Global gene expression analysis in human uterine epithelial cells defines new targets of glucocorticoid and estradiol antagonism. *Biology of Reproduction*, 89(3):66–1.

Williamson, M., Garg, R., and Wells, C. M. (2020). PlexinB1 promotes nuclear translocation of the glucocorticoid receptor. *Cells*, 9(1):3.

Wolff, N. C., McKay, R. M., and Brugarolas, J. (2014). REDD1/DDIT4-independent mTORC1 inhibition and apoptosis by glucocorticoids in thymocytes. *Molecular Cancer Research*, 12(6):867–877.

Xing, X., Zhao, Z., and Liu, J. S. (2019). Controlling false discovery rate using Gaussian mirrors. *arXiv:1911.09761*.

Zhang, C. H. and Zhang, S. S. (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.

Zhang, X., Zhang, F., Guo, L., Wang, Y., Zhang, P., Wang, R., Zhang, N., and Chen, R. (2013). Interactome analysis reveals that C1QBP (complement component 1, q subcomponent binding protein) is associated with cancer cell chemotaxis and metastasis. *Molecular & Cellular Proteomics*, 12(11):3199–3209.

Zhao, Q., Sur, P., and Candès, E. J. (2020). The asymptotic distribution of the MLE in high-dimensional logistic models: arbitrary covariance. *arXiv:2001.09351*.

Zhao, W., Wang, D., Liu, C., and Zhao, X. (2016). G-protein-coupled receptor kinase 2 terminates G-protein-coupled receptor function in steroid hormone 20-hydroxyecdysone signaling. *Scientific Reports*, 6(1):1–13.

Zhou, A., Wei, Z., Stanley, P. L. D., Read, R. J., Stein, P. E., and Carrell, R. W. (2008). The S-to-R transition of corticosteroid-binding globulin and the mechanism of hormone release. *Journal of Molecular Biology*, 380(1):244–251.

Zilliacus, J., Holter, E., Wakui, H., Tazawa, H., Treuter, E., and Gustafssone, J. (2001). Regulation of glucocorticoid receptor activity by 14–3-3-dependent intracellular relocalization of the corepressor RIP140. *Molecular Endocrinology*, 15(4):501–511.

Zorzatto, C., Machado, J. P. B., Lopes, K. V. G., Nascimento, K. J. T., Pereira, W. A., Brustolini, O. J. B., Reis, P. A. B., Calil, I. P., Deguchi, M., and Sachetto-Martins, G. (2015). NIK1-mediated translation suppression functions as a plant antiviral immunity mechanism. *Nature*, 520(7549):679–682.

Appendix

A Proof

Let $Q(t) = 1 - \Phi(t)$, where Φ is the CDF of the standard normal distribution. Let $\phi(t)$ be the probability density function of the standard normal distribution. We consider a general form of the mirror statistic defined in Equation (4) with $f(u, v)$ satisfying Condition 2.2. For $t \in \mathbb{R}$, let $H(t) = \mathbb{P}(\text{sign}(Z_1 Z_2) f(Z_1, Z_2) > t)$, where Z_1 and Z_2 are independent, following the standard normal distribution. For any $t > 0$ and $v \geq 0$, let

$$\mathcal{I}_t(v) = \inf\{u \geq 0 : f(u, v) > t\} \tag{34}$$

with the convention $\inf \varnothing = +\infty$.

A.1 Proof of Proposition 2.1

Let $r = \lim p_1/p_0$. Without loss of generality, we assume the designated FDR control level $q \in (0, 1)$ satisfying $rq/(1-q) < 1$, otherwise selecting all the features would achieve an asymptotic FDR control.

Denote $f^{\text{opt}}(u, v)$ as the optimal choice, and denote $\widehat{S}^{\text{opt}}$ as the optimal selection result that enjoys an asymptotic FDR control. By the law of large number, we have

$$\lim_{p\to\infty} \frac{\#\{j : j \in S_0, j \in \widehat{S}^{\text{opt}}\}}{\#\{j : j \in \widehat{S}^{\text{opt}}\}} = \frac{\mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_0)}{\mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_0) + r\mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_1)} \leq q, \tag{35}$$

in which the numerator is essentially the type-I error. More precisely,

$$\begin{aligned} \mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_0) &= \mathbb{P}(\text{sign}(Z_1 Z_2) f^{\text{opt}}(|Z_1|, |Z_2|) > t^{\text{opt}}), \\ \mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_1) &= \mathbb{P}(\text{sign}(Z_3 Z_4) f^{\text{opt}}(|Z_3|, |Z_4|) > t^{\text{opt}}), \end{aligned} \tag{36}$$

where Z_1, Z_2 follow $N(0, 1)$, Z_3, Z_4 follow $N(\omega, 1)$, and all of them are independent. $t^{\text{opt}} > 0$ is the cutoff that maximizes the power $\mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_1)$, under the constraint that Equation (35) is satisfied.

We now consider testing whether X_j is a null feature, with the significance level α specified as below,

$$\alpha = \frac{rq}{1-q} \mathbb{P}(j \in \widehat{S}^{\text{opt}} \mid j \in S_1) < 1. \tag{37}$$

We note that we essentially have two observations $T_j^{(1)}$ and $T_j^{(2)}$, which independently follow $N(0, 1)$ or $N(\omega, 1)$ if X_j is a null feature or a relevant feature, respectively. By Equation (35), the test which rejects the null hypothesis (i.e., $j \in \widehat{S}^{\text{opt}}$) if

$$\text{sign}(T_j^{(1)} T_j^{(2)}) f^{\text{opt}}(|T_j^{(1)}|, |T_j^{(2)}|) > t^{\text{opt}} \tag{38}$$

achieves the significance level α.

By the Neymann-Pearson Lemma, the likelihood ratio test is the most powerful test, which rejects the null hypothesis if the following likelihood ratio (LR) is large enough,

$$\mathrm{LR} = \frac{\phi(T_j^{(1)} - \omega)\phi(T_j^{(2)} - \omega)}{\phi(T_j^{(1)})\phi(T_j^{(1)})} \approx \frac{\phi(|T_j^{(1)}| - \omega)\phi(|T_j^{(2)}| - \omega)}{\phi(|T_j^{(1)}|)\phi(|T_j^{(1)}|)} \tag{39}$$

The approximation shall be reasonably accurate when $\omega > 0$ is sufficiently large. A simple calculation yields the following rejection rule

$$|T_j^{(1)}| + |T_j^{(2)}| > t^{\mathrm{lik}}, \tag{40}$$

in which t^{lik} is a properly chosen cutoff so that the likelihood ratio test achieves the significance level α.

We then consider the following rejection rule

$$\mathrm{sign}(T_j^{(1)}T_j^{(2)}) f^{\mathrm{lik}}(|T_j^{(1)}|, |T_j^{(2)}|) > t^{\mathrm{lik}}, \tag{41}$$

in which $f^{\mathrm{lik}}(u, v) = u + v$. Denote $\{j \in \widehat{S}^{\mathrm{lik}}\}$ as the event of rejecting the null hypothesis. We note that the type-I error is upper bounded by α since

$$\mathbb{P}(\{j \in \widehat{S}^{\mathrm{lik}}\} \mid j \in S_0) \le \mathbb{P}(|T_j^{(1)}| + |T_j^{(2)}| > t^{\mathrm{lik}} \mid j \in S_0) \le \alpha. \tag{42}$$

In addition, since the likelihood ratio test is the optimal test, we have

$$\begin{aligned} \mathbb{P}(j \in \widehat{S}^{\mathrm{lik}} \mid j \in S_1) &\approx \mathbb{P}(|T_j^{(1)}| + |T_j^{(2)}| > t^{\mathrm{lik}} \mid j \in S_1) \\ &\ge \mathbb{P}(j \in \widehat{S}^{\mathrm{opt}} \mid j \in S_1). \end{aligned} \tag{43}$$

when $\omega > 0$ is sufficiently large. Combining Equations (42) and (43), we show that the selection set $\widehat{S}^{\mathrm{lik}}$ enjoys an asymptotic FDR control since

$$\lim_{p\to\infty} \frac{\#\{j : j \in S_0, j \in \widehat{S}^{\mathrm{lik}}\}}{\#\{j : j \in \widehat{S}^{\mathrm{lik}}\}} = \frac{\mathbb{P}(j \in \widehat{S}^{\mathrm{lik}} \mid j \in S_0)}{\mathbb{P}(j \in \widehat{S}^{\mathrm{lik}} \mid j \in S_0) + r\mathbb{P}(j \in \widehat{S}^{\mathrm{lik}} \mid j \in S_1)} \le q. \tag{44}$$

Since f^{opt} is optimal, by Equations (44) and (43), f^{lik} is also optimal. This concludes the proof of Proposition 2.1.

A.2 Proof of Proposition 3.1

A.2.1 Technical Lemmas

Lemma A.1. Consider the case $\Sigma = I_p$. As $n, p \to \infty$, we have

$$\sigma_n \xrightarrow{p} \sqrt{\kappa}\sigma_\star, \quad \alpha_n \xrightarrow{p} \alpha_\star, \tag{45}$$

in which $(\sigma_n,\ \alpha_n)$ is defined in Equation (13), and $(\sigma_\star, \alpha_\star)$ is the unique optimizer of the following optimization problem,

$$\begin{aligned} \min_{\sigma,\delta>0,\alpha\in\mathbb{R}} \max_{r>0} \Big\{ r\kappa\sigma + \frac{r}{2\delta} - \frac{1}{2r\delta}\mathbb{E}(y_i^2) - \alpha\mathbb{E}(y_i x_i^{\intercal}\beta^\star) \\ + \mathbb{E}\Big[M_\rho\Big(\alpha\gamma Z_1 + \sigma\sqrt{\kappa}Z_2 + \frac{y_i}{r\delta}, \frac{1}{r\delta}\Big)\Big]\Big\}. \end{aligned} \tag{46}$$

Z_1, Z_2 are independent (also independent to everything else) random variables following the standard normal distribution, and

$$M_\rho(v, t) = \min_{x\in\mathbb{R}} \left\{ \rho(x) + \frac{1}{2t}(x - v)^2 \right\}. \tag{47}$$

Proof of Lemma A.1. The proof heavily relies on Thrampoulidis et al. (2018) and Salehi et al. (2019), in which the key technical tool is the Convex Gaussian Min-max Theorem(CGMT). We introduce a new variable $u = X\beta$ and rewrite the optimization problem (11) as

$$\min_{\beta\in\mathbb{R}^p, u\in\mathbb{R}^n} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u \quad s.t. \quad u = X\beta. \tag{48}$$

Based on the method of Lagrange multipliers, the above optimization problem is equivalent to a min-max problem specified as below,

$$\text{(PO)} \quad \min_{\beta\in\mathbb{R}^p, u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^n} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u + \frac{1}{n}v^\intercal(u - X\beta). \tag{49}$$

We refer to this min-max optimization problem as the primary optimization (PO) problem henceforth.

We proceed to associate the PO with an auxiliary optimization (AO) problem using CGMT. Similar as in Salehi et al. (2019), We decompose $\beta = P\beta + P^\perp\beta$, in which P is the projection operator projecting onto the column space spanned by $\beta^\star$, and rewrite the PO as below,

$$\min_{\beta\in\mathbb{R}^p, u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^n} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u + \frac{1}{n}v^\intercal u - \frac{1}{n}v^\intercal XP\beta - \frac{1}{n}v^\intercal XP^\perp\beta. \tag{50}$$

Before we apply CGMT, we remark that we can simply assume the feasible sets of v, u and β are convex and compact, following Lemma A.1 and Lemma A.2 in Thrampoulidis et al. (2018). For simplicity, we omit the details here and write the AO as:

$$\text{(AO)} \quad \min_{\beta\in\mathbb{R}^p, u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^n} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u + \frac{1}{n}v^\intercal(u - XP\beta) - \frac{1}{n}(v^\intercal h||P^\perp\beta|| + ||v||g^\intercal P^\perp\beta), \tag{51}$$

in which h and g are independent, following $N(0, I_n)$ and $N(0, I_p)$, respectively.

We proceed to simplify the AO. Let $r = ||v||/\sqrt{n}$. Maximizing with respect to the direction of v, the AO can be written as

$$\min_{\beta\in\mathbb{R}^p, u\in\mathbb{R}^n} \max_{r\geq 0} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u - \frac{r}{\sqrt{n}}g^\intercal P^\perp\beta + r\left|\left|\frac{u}{\sqrt{n}} - \frac{XP\beta}{\sqrt{n}} - \frac{||P^\perp\beta||h}{\sqrt{n}}\right|\right|. \tag{52}$$

By Lemma A.3 in Thrampoulidis et al. (2018), we swap the order of min-max in the optimization problem (52). Let $\sigma = ||P^\perp\beta||/\sqrt{\kappa}$, $\alpha = \langle\beta, \beta^\star\rangle / ||\beta^\star||^2$, then $P\beta = \alpha\beta^\star$. Optimizing with respect to the direction of $P^\perp\beta$, we can further simplify the optimization problem (52) as

$$\max_{r\geq 0} \min_{u\in\mathbb{R}^n, \sigma\geq 0, \alpha\in\mathbb{R}} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u + \frac{r\sqrt{p}}{n}||g||\sigma + r\left|\left|\frac{u}{\sqrt{n}} - \frac{\alpha X\beta^\star}{\sqrt{n}} - \frac{\sqrt{\kappa}\sigma h}{\sqrt{n}}\right|\right|. \tag{53}$$

Using the square-root trick, i.e., $\sqrt{x} = \inf_{\delta>0}\left\{\frac{\delta}{2} + \frac{x}{2\delta}\right\}$, we obtain

$$\max_{r\geq 0} \min_{\substack{u\in\mathbb{R}^n, \sigma\geq 0\\ \alpha\in\mathbb{R},\ \delta>0}} \frac{1}{n}1^\intercal\rho(u) - \frac{1}{n}y^\intercal u + \frac{r\sqrt{p}}{n}||g||\sigma + \frac{\delta r}{2}\left|\left|\frac{u}{\sqrt{n}} - \frac{\alpha X\beta^\star}{\sqrt{n}} - \frac{\sqrt{\kappa}\sigma h}{\sqrt{n}}\right|\right|^2 + \frac{r}{2\delta}. \tag{54}$$

We proceed to optimize with respect to u. Based on completion of square, we have

$$\begin{aligned} &-\frac{1}{n}y^\intercal u + \frac{\delta r}{2}\left|\left|\frac{u}{\sqrt{n}} - \frac{\alpha X\beta^\star}{\sqrt{n}} - \frac{\sqrt{\kappa}\sigma h}{\sqrt{n}}\right|\right|^2 \\ &\qquad = \frac{\delta r}{2n}\left|\left|u - \alpha X\beta^\star - \sqrt{\kappa}\sigma h - \frac{y}{\delta r}\right|\right|^2 - \frac{||y||^2}{2n\delta r} - \frac{\alpha}{n}y^\intercal X\beta^\star - \frac{\sqrt{\kappa}}{n}\sigma y^\intercal h. \end{aligned} \tag{55}$$

Optimizing with respect to the direction of u, the AO can be simplified as below,

$$\max_{r\geq 0}\min_{\sigma\geq 0,\alpha\in\mathbb{R},\delta>0}\mathcal{R}_n(\sigma,\alpha,\delta,r) := \frac{r\sqrt{p}}{n}||g||\sigma + \frac{r}{2\delta} - \frac{||y||^2}{2n\delta r} - \frac{\alpha}{n}y^\intercal X\beta^\star - \frac{\sqrt{\kappa}}{n}\sigma y^\intercal h + \frac{1}{n}M_\rho\Big(\alpha X\beta^\star + \sigma\sqrt{\kappa}h + \frac{y}{\delta r}, \frac{1}{\delta r}\Big), \tag{56}$$

in which the function M_ρ is applied in an element-wise fashion. We now reduce the original vector optimization problem into a scalar optimization problem. Since the objective function is convex in σ, α, δ and concave in r, we can swap the order of min and max.

We proceed to further simplify the AO by invoking asymptotics. By the law of large numbers, we have the following simple results,

$$\begin{aligned}
&\text{(a)}\quad \frac{1}{n}||y||^2 \xrightarrow{p} \mathbb{E}(y^2);\\
&\text{(b)}\quad \frac{1}{n}y^\intercal X\beta^\star \xrightarrow{p} \mathbb{E}(y_i X_i^\intercal \beta^\star);\\
&\text{(c)}\quad \frac{1}{n}M_\rho\Big(\alpha X\beta^\star + \sigma\sqrt{\kappa}h + \frac{y}{\delta r}, \frac{1}{\delta r}\Big) \xrightarrow{p} \mathbb{E}\Big[M_\rho\Big(\alpha\gamma Z_1 + \sigma\sqrt{\kappa}Z_2 + \frac{y}{\delta r}, \frac{1}{\delta r}\Big)\Big],\\
&\text{(d)}\quad \frac{1}{n}y^\intercal h = \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n}}y^\intercal h = O_p(1/\sqrt{n}),\\
&\text{(e)}\quad \frac{\sqrt{p}}{n}||g|| = \frac{p}{n}\frac{||g||}{\sqrt{p}} \xrightarrow{p} \frac{p}{n} = \kappa.
\end{aligned} \tag{57}$$

Based upon these facts, $\mathcal{R}_n$ converges in probability to

$$D(\sigma,\alpha,\delta,r) := r\kappa\sigma + \frac{r}{2\delta} - \frac{1}{2\delta r}\mathbb{E}(y^2) - \alpha\mathbb{E}(yX^\intercal\beta^\star) + \mathbb{E}\Big[M_\rho\Big(\alpha\gamma Z_1 + \sigma\sqrt{\kappa}Z_2 + \frac{y}{\delta r}, \frac{1}{\delta r}\Big)\Big]. \tag{58}$$

We note that the objective function $D(\sigma,\alpha,\delta,r)$ is convex in σ,α,δ and concave in r, since the convexity is preserved through point-wise limit. The covergence of the optimizers can be established based on the same arguments as in Lemma B.1 and Lemma A.5 in Thrampoulidis et al. (2018). This completes the proof of Lemma A.1.

A.2.2 Proof of Proposition 3.1

The proof of Proposition 3.1 is essentially the same as the proof of Theorem 3.1 in Zhao et al. (2020). Denote $\Sigma = LL^\intercal$ as the Cholesky decomposition of the covariance matrix Σ. Let

$$\theta^\star = L^\intercal\beta^\star, \quad \widehat{\theta} = L^\intercal\widehat{\beta}, \tag{59}$$

in which $\widehat{\beta}$ is the MLE of the true regression coefficient $\beta^\star$. By Proposition 2.1 in Zhao et al. (2020), $\widehat{\theta}$ has the same distribution as the MLE of the underlying GLM with true regression coefficient $\theta^\star$ and features drawn i.i.d from $N(0, I_p)$. We note that it is sufficient to consider the case $\Sigma = I_p$, as the general result follows from the following relationship,

$$\tau_j\frac{\widehat{\beta}_j - \alpha_\star\beta_j^\star}{\sigma_\star} = \frac{\widehat{\theta}_j - \alpha_\star\theta_j^\star}{\sigma_\star}. \tag{60}$$

For $j\in[p]$, we have the following decomposition,

$$\frac{\sqrt{n}(\widehat{\theta}_j - \alpha_\star\theta_j^\star)}{\sigma_\star} = \frac{\sqrt{n}(\widehat{\theta}_j - \alpha_n\theta_j^\star)}{\sigma_n}\frac{\sigma_n}{\sigma_\star} + \frac{\sqrt{n}(\alpha_n - \alpha_\star)\theta_j^\star}{\sigma_\star}. \tag{61}$$

Using the same arguments as in the proof of Theorem 3.1 in Zhao et al. (2020), we can show that the first term asymptotically behaves as the standard normal distribution, whereas the second term asymptotically vanishes. This completes the proof of Proposition 3.1.

A.3 Proof of Proposition 3.2

In addition to the notations introduced in Equation (59), we require the following notations for the ease of presentation. Let $\langle u, v\rangle_\Sigma = u^\intercal \Sigma v$ for any $u, v \in \mathbb{R}^p$, and $||u||_\Sigma^2 = u^\intercal \Sigma u$. We introduce the following random vector,

$$\xi = W - \frac{1}{\gamma_n^2}\langle W, \beta^\star\rangle_\Sigma \beta^\star, \tag{62}$$

in which W follows $N(0, \Theta)$, independent to everything else. Recall that $\gamma_n^2 = \mathrm{Var}(x_i^\intercal \beta^\star)$ calibrates the signal strength, and we assume $\gamma_n \to \gamma$ as $n, p \to \infty$. We define

$$\widetilde{V}_j = \frac{\sqrt{p} V_j}{||\xi||_\Sigma} \quad \text{with} \quad V_j = \tau_j W_j, \tag{63}$$

for $j \in [p]$ where $\tau_j^2 = 1/\Theta_{jj}$. The random vector V follows $N(0, R)$ with $R_{ij} = \tau_i \tau_j \Theta_{ij}$. With a bit abuse of notation, we denote $\kappa = 2p/n \in (0, 1)$ as the ratio of the dimension over the sample size after data splitting.

We define the normalized MLE T_j as well as its approximation $\widetilde{T}_j$ as below,

$$T_j = \frac{\sqrt{n}\widehat{\tau}_j \widehat{\beta}_j}{\sigma_\star}, \quad \widetilde{T}_j = \frac{\sqrt{n\kappa}\tau_j \widehat{\beta}_j}{\sigma_n} \tag{64}$$

for $j \in [p]$, in which $\sigma_\star$ is the unique optimizer of the optimization problem defined in (46). Without changing the selection result obtained via Algorithm 1, we multiply the normalized MLE defined in Equation (14) by a constant factor so that it follows the standard normal distribution asymptotically. Let $\widetilde{M}_j$ be the corresponding approximated mirror statistic constructed based upon $\widetilde{T}_j^{(1)}$ and $\widetilde{T}_j^{(2)}$.

A.3.1 Technical lemmas

Lemma A.2. $\widetilde{T}_{S_0} \overset{d}{=} \widetilde{V}_{S_0}$.

Proof of Lemma A.2. We note that a simple calculation yields

$$L^{-\intercal} P_{\theta^\star}^\perp Z \overset{d}{=} \xi, \quad ||P_{\theta^\star}^\perp Z||_2 = ||\xi||_\Sigma, \tag{65}$$

in which Z following $N(0, I_p)$ is independent to everything else. Let Λ_{S_0} be a $p_0 \times p_0$ diagonal matrix, of which the diagonal elements are τ_j for $j \in S_0$. By the stochastic representation (Lemma 2.1, Proposition A.1) in Zhao et al. (2020), we have

$$\begin{aligned} \widetilde{T}_{S_0} &= \frac{\sqrt{p}\Lambda_{S_0}\widehat{\beta}_{S_0}}{\sigma_n} = \frac{\sqrt{p}\Lambda_{S_0}(L^{-\intercal}\widehat{\theta})_{S_0}}{\sigma_n} \\ &\overset{d}{=} \frac{\sqrt{p}\Lambda_{S_0}(L^{-\intercal}P_{\theta^\star}^\perp Z)_{S_0}}{||P_{\theta^\star}^\perp Z||_2} \overset{d}{=} \frac{\sqrt{p}\Lambda_{S_0}\xi_{S_0}}{||\xi||_\Sigma} \overset{d}{=} \widetilde{V}_{S_0}. \end{aligned} \tag{66}$$

This completes the proof of Lemma A.2.

Lemma A.3. Under Assumption 3.1, there exists a constant $c > 0$ such that for $\epsilon \in (0, 1)$, we have

$$\mathbb{P}\left(\max_{j\in[p]} \left|\widehat{\tau}_j^2/\tau_j^2 - 1\right| > \epsilon\right) \le p\exp(-c(n/2 - p + 1)\epsilon^2). \tag{67}$$

Proof of Lemma A.3. By Assumption 3.1, we have

$$\begin{aligned} \min_{j\in[p]} \tau_j^2 &= 1/\max_{j\in[p]} \Theta_{jj} \ge 1/\sigma_{\max}(\Theta) = \sigma_{\min}(\Sigma) \ge 1/C > 0, \\ \max_{j\in[p]} \tau_j^2 &= 1/\min_{j\in[p]} \Theta_{jj} \le 1/\sigma_{\min}(\Theta) = \sigma_{\max}(\Sigma) \le C < \infty. \end{aligned} \tag{68}$$

Thus, it is sufficient for us to consider $\max_{j\in[p]}\left|\widehat{\tau}_j^2-\tau_j^2\right|$. Since $\mathrm{RSS}_j\sim\tau_j^2\chi^2_{n/2-p+1}$, by the union bound and a Bernstein-type inequality, for any $\epsilon\in(0,1)$, we have

$$\begin{aligned}\mathbb{P}\left(\max_{j\in[p]}\left|\widehat{\tau}_j^2-\tau_j^2\right|>\epsilon\right)&\leq\sum_{j=1}^{p}\mathbb{P}\left(\left|\widehat{\tau}_j^2-\tau_j^2\right|>\epsilon\right)\\&\leq p\exp(-c(n/2-p+1)\epsilon^2)\end{aligned}\tag{69}$$

for some constant $c>0$.

Lemma A.4. Under Assumption 3.1, as $n,p\to\infty$, we have

$$\sup_{t\in\mathbb{R},j\in S_0}|\mathbb{P}(T_j>t)-Q(t)|\longrightarrow 0\tag{70}$$

Proof of Lemma A.4. Without loss of generality, we assume $t>0$. For $\epsilon\in(0,1)$, we condition on the event $E_1\cap E_2$, in which

$$\begin{aligned}E_1&=\{|\sigma_n/\sqrt{\kappa}-\sigma_\star|<\epsilon\},\\E_2&=\left\{\max_{j\in[p]}|\widehat{\tau}_j/\tau_j-1|<\epsilon\right\}.\end{aligned}\tag{71}$$

For large enough n and p, E_1 and E_2 hold with high probability according to Lemma A.1 and Lemma A.3, respectively. Conditioning on the events E_1 and E_2, we have

$$\begin{aligned}\max_{j\in[p]}|T_j/\widetilde{T}_j-1|&\leq\max_{j\in[p]}|\widehat{\tau}_j/\tau_j||\sigma_n/\sqrt{\kappa}-\sigma_\star|/\sigma_\star+\max_{j\in[p]}|\widehat{\tau}_j/\tau_j-1|\\&\leq(\epsilon+1)\epsilon/\sigma_\star+\epsilon.\end{aligned}\tag{72}$$

Consequently, we have

$$\max_{j\in[p]}|T_j/\widetilde{T}_j-1|\xrightarrow{p}0.\tag{73}$$

We proceed to show that

$$\max_{j\in[p]}\left|\widetilde{V}_j/V_j-1\right|=\left|\frac{\sqrt{p}}{||\xi||_\Sigma}-1\right|\xrightarrow{p}0,\tag{74}$$

in which by definition, V_j follows the standard normal distribution. Indeed, by the definition of ξ in Equation (62), we have

$$\frac{||\xi||_\Sigma^2}{p}=\frac{||W||_\Sigma^2}{p}-\frac{1}{p}\langle W,\frac{\beta^\star}{\gamma_n}\rangle_\Sigma^2.\tag{75}$$

The second term converges to 0 in probability, since $\langle W,\beta^\star/\gamma_n\rangle_\Sigma$ follows the standard normal distribution. For the first term, we have

$$\frac{||W||_\Sigma^2}{p}=\frac{1}{p}W^\intercal\Sigma W=\frac{Z^\intercal Z}{p}\xrightarrow{p}1.\tag{76}$$

It follows that

$$\begin{aligned}\sup_{t>0,\ j\in S_0}|\mathbb{P}(T_j>t)-Q(t)|\leq&\sup_{t>0,\ j\in S_0}\left|\mathbb{P}(T_j>t)-\mathbb{P}(\widetilde{T}_j>t)\right|\\&+\sup_{t>0,\ j\in S_0}\left|\mathbb{P}(\widetilde{T}_j>t)-Q(t)\right|\\=&\sup_{t>0,\ j\in S_0}\left|\mathbb{P}(T_j>t)-\mathbb{P}(\widetilde{T}_j>t)\right|\\&+\sup_{t>0,\ j\in S_0}\left|\mathbb{P}(\widetilde{V}_j>t)-\mathbb{P}(V_j>t)\right|\\\xrightarrow{p}&\ 0.\end{aligned}\tag{77}$$

The equality follows from Lemma A.2. The convergence in the last line follows from Equations (73) and (74) (detailed arguments can be found in Equations (82) and (83)). This completes the proof of Lemma A.4.

Lemma A.5. If Equation (70) holds, as $n, p \to \infty$, we have

$$\sup_{t\in\mathbb{R},\ j\in S_0} |\mathbb{P}(M_j > t) - H(t)| \longrightarrow 0. \tag{78}$$

Proof of Lemma A.5. Without loss of generality, we assume $t > 0$. Let $Z^{(1)}$ and $Z^{(2)}$ be two independent random variables following the standard normal distribution, which are also independent to $T_j^{(1)}$ and $T_j^{(2)}$. For any $\epsilon > 0$, for large enough n and p, we have

$$\begin{aligned}
\mathbb{P}(M_j > t) &= \mathbb{P}\Big(T_j^{(2)} > I_t\big(T_j^{(1)}\big),\ T_j^{(1)} > 0\Big) + \mathbb{P}\Big(T_j^{(2)} < -I_t\big(T_j^{(1)}\big),\ T_j^{(1)} < 0\Big) \\
&\le \mathbb{P}\Big(Z^{(2)} > I_t\big(T_j^{(1)}\big),\ T_j^{(1)} > 0\Big) + \mathbb{P}\Big(Z^{(2)} < -I_t\big(T_j^{(1)}\big),\ T_j^{(1)} < 0\Big) + \epsilon \\
&= \mathbb{P}\Big(\mathrm{sign}\big(T_j^{(1)} Z^{(2)}\big) f\big(T_j^{(1)}, Z^{(2)}\big) > t\Big) + \epsilon \\
&= \mathbb{P}\Big(T_j^{(1)} > I_t\big(Z^{(2)}\big),\ Z^{(2)} > 0\Big) + \mathbb{P}\Big(T_j^{(1)} < -I_t\big(Z^{(2)}\big),\ Z^{(2)} < 0\Big) + \epsilon \\
&\le \mathbb{P}\Big(Z^{(1)} > I_t\big(Z^{(2)}\big),\ Z^{(2)} > 0\Big) + \mathbb{P}\Big(Z^{(1)} < -I_t\big(Z^{(2)}\big),\ Z^{(2)} < 0\Big) + 2\epsilon \\
&= H(t) + 2\epsilon,
\end{aligned} \tag{79}$$

where all the equalities follow from the fact that $f(u, v)$ is monotonically increasing with respect to $|u|$ and $|v|$, and the two inequalities follow from Equation (70) as well as the independence between the random variables. Similarly, we can show that $\mathbb{P}(M_j > t) \ge H(t) - 2\epsilon$ for large enough n and p. This completes the proof of Lemma A.5.

Lemma A.6. Under Assumption 3.1, as $n, p \to \infty$, we have

$$\sup_{i,j\in S_0,\ t_1,t_2\in\mathbb{R}} |\mathbb{P}(T_i > t_1, T_j > t_2) - \mathbb{P}(V_i > t_1, V_j > t_2)| \longrightarrow 0. \tag{80}$$

Proof of Lemma A.6. Without loss of generality, we assume $t_1 > 0$ and $t_2 > 0$. For any given $\epsilon > 0$, denote

$$E_1 = \Big\{ \max_{k\in[p]} |T_k/\widetilde{T}_k - 1| \le \epsilon \Big\}, \quad E_2 = \Big\{ \max_{k\in[p]} |\widetilde{V}_k/V_k - 1| \le \epsilon \Big\}. \tag{81}$$

By Equations (73) and (74), for large enough n and p, we have $\mathbb{P}(E_1) \ge 1 - \epsilon$ and $\mathbb{P}(E_2) \ge 1 - \epsilon$. For $i, j \in S_0$, we have

$$\begin{aligned}
\mathbb{P}(T_i > t_1, T_j > t_2) &\ge \mathbb{P}(T_i > t_1, T_j > t_2 \mid E_1)\mathbb{P}(E_1) \\
&\ge \mathbb{P}\Big(\widetilde{T}_i > \frac{t_1}{1-\epsilon}, \widetilde{T}_j > \frac{t_2}{1-\epsilon} \mid E_1\Big)\mathbb{P}(E_1) \\
&\ge \mathbb{P}\Big(\widetilde{T}_i > \frac{t_1}{1-\epsilon}, \widetilde{T}_j > \frac{t_2}{1-\epsilon}\Big) - \epsilon \\
&= \mathbb{P}\Big(\widetilde{V}_i > \frac{t_1}{1-\epsilon}, \widetilde{V}_j > \frac{t_2}{1-\epsilon}\Big) - \epsilon \\
&\ge \mathbb{P}\Big(\widetilde{V}_i > \frac{t_1}{1-\epsilon}, \widetilde{V}_j > \frac{t_2}{1-\epsilon} \mid E_2\Big)\mathbb{P}(E_2) - \epsilon \\
&\ge \mathbb{P}\Big(V_i > \frac{t_1}{(1-\epsilon)^2}, V_j > \frac{t_2}{(1-\epsilon)^2} \mid E_2\Big)\mathbb{P}(E_2) - \epsilon \\
&\ge \mathbb{P}\Big(V_i > \frac{t_1}{(1-\epsilon)^2}, V_j > \frac{t_2}{(1-\epsilon)^2}\Big) - 2\epsilon,
\end{aligned} \tag{82}$$

in which the equality in the fourth line follows from Lemma A.2. Similarly, we have

$$\mathbb{P}(T_i > t_1, T_j > t_2) \leq \mathbb{P}(V_i > (1-\epsilon)^2 t_1, V_j > (1-\epsilon)^2 t_2) + 2\epsilon. \tag{83}$$

Combining Equations (82) and (83) leads to the claim in Lemma A.6.

Corollary A.1. Under Assumption 3.1, for any Borel set B in $\mathbb{R}^2$, as $n, p \to \infty$, we have

$$\sup_{i,j \in S_0 \in \mathbb{R}} |\mathbb{P}((T_i, T_j) \in B) - \mathbb{P}((V_i, V_j) \in B)| \longrightarrow 0. \tag{84}$$

Proof of Corollary A.1. The proof follows immediately from Lemma A.6 and Example 2.3 in Billingsley (2013).

Lemma A.7. Under Assumption 3.1, as $n, p \to \infty$, we have

$$\sup_{t \in \mathbb{R}} \mathrm{Var}\left(\frac{1}{p_0} \sum_{j \in S_0} \mathbb{1}(M_j > t)\right) \leq \frac{1}{4p_0} + O(||R_{S_0}||_1) + o(1), \tag{85}$$

in which $||R_{S_0}||_1 = \sum_{i,j \in S_0} R_{ij}/p_0^2$.

Proof of Lemma A.7. Without loss of generality, we assume $t > 0$. We have

$$\begin{aligned} \sup_{t \in \mathbb{R}} \mathrm{Var}\left(\frac{1}{p_0} \sum_{j \in S_0} \mathbb{1}(M_j > t)\right) \leq & \frac{1}{p_0^2} \sum_{j \in S_0} \sup_{t \in \mathbb{R}} \mathrm{Var}(\mathbb{1}(M_j > t)) \\ & + \frac{1}{p_0^2} \sum_{i \neq j \in S_0} \sup_{t \in \mathbb{R}} \mathrm{Cov}(\mathbb{1}(M_i > t), \mathbb{1}(M_j > t)). \end{aligned} \tag{86}$$

Using the Cauchy-Schwartz inequality, the first term is upper bounded by $1/(4p_0)$. For the second term, since

$$\begin{aligned} \mathrm{Cov}(\mathbb{1}(M_i > t), \mathbb{1}(M_j > t)) \leq & \left|\mathbb{P}(M_i > t, M_j > t) - H^2(t)\right| \\ & + \left|\mathbb{P}(M_i > t)\mathbb{P}(M_j > t) - H^2(t)\right|, \end{aligned} \tag{87}$$

by Lemma A.5, it is sufficient for us to show that for $i, j \in S_0$ and $i \neq j$,

$$\sup_{t \in \mathbb{R}} \left|\mathbb{P}(M_i > t, M_j > t) - H^2(t)\right| \leq O(|R_{ij}|) + o(1). \tag{88}$$

By Condition 2.2, we have the following decomposition based on the signs of $T_i^{(1)}$ and $T_j^{(1)}$,

$$\begin{aligned} \mathbb{P}(M_i > t, M_j > t) = & \mathbb{P}\left(T_i^{(2)} > I_t(T_i^{(1)}),\ \ T_j^{(2)} > I_t(T_j^{(1)}),\ \ T_i^{(1)} > 0, T_j^{(1)} > 0\right) \\ & + \mathbb{P}\left(T_i^{(2)} > I_t(T_i^{(1)}),\ \ T_j^{(2)} < -I_t(T_j^{(1)}), T_i^{(1)} > 0, T_j^{(1)} < 0\right) \\ & + \mathbb{P}\left(T_i^{(2)} < -I_t(T_i^{(1)}), T_j^{(2)} > I_t(T_j^{(1)}),\ \ T_i^{(1)} < 0, T_j^{(1)} > 0\right) \\ & + \mathbb{P}\left(T_i^{(2)} < -I_t(T_i^{(1)}), T_j^{(2)} < -I_t(T_j^{(1)}), T_i^{(1)} < 0, T_j^{(1)} < 0\right) \\ & := I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{89}$$

Denote Φ_r and ϕ_r as the CDF and pdf of the bivariate normal distribution with variances 1 and correlation r, respectively. Let $\phi^{(n)}$ be the n-th derivative of ϕ. Recall the Mehler's identity (Kotz et al., 2000), that is, for any $t_1, t_2 \in \mathbb{R}$,

$$\Phi_r(t_1, t_2) = \Phi(t_1)\Phi(t_2) + \sum_{n=1}^{\infty} \frac{r^n}{n!} \phi^{(n-1)}(t_1)\phi^{(n-1)}(t_2). \tag{90}$$

For I_1, we have the following upper bound,

$$\begin{aligned} I_1 &= \mathbb{E}\left[\mathbb{P}\left(T_i^{(2)} > I_t(x),\ T_j^{(2)} > I_t(y)\right) \mid T_i^{(1)} = x > 0,\ T_j^{(1)} = y > 0\right] \\ &= \mathbb{E}\left[\mathbb{P}\left(V_i^{(2)} > I_t(x),\ V_j^{(2)} > I_t(y)\right) \mid T_i^{(1)} = x > 0,\ T_j^{(1)} = y > 0\right] + o(1) \\ &\leq \mathbb{E}\left[Q(I_t(x))Q(I_t(y)) \mid T_i^{(1)} = x > 0,\ T_j^{(1)} = y > 0\right] + O(|R_{ij}|) + o(1). \end{aligned} \tag{91}$$

The second line follows from Lemma A.6. The third line follows from the Mehler's identity and Lemma 1 in Azriel and Schwartzman (2015), i.e.,

$$\sum_{n=1}^{\infty} \frac{\left[\sup_{t\in\mathbb{R}} \phi^{(n-1)}(t)\right]^2}{n!} < \infty. \tag{92}$$

Similarly, we can upper bound I_2, I_3 and I_4. Combining the four upper bounds together, we obtain an upper bound on $\mathbb{P}(M_i > t, M_j > t)$ specified as below,

$$\mathbb{P}\left(\text{sign}(Z_i^{(2)}T_i^{(1)})f(Z_i^{(2)}, T_i^{(1)}) > t, \text{sign}(Z_j^{(2)}T_j^{(1)})f(Z_j^{(2)}, T_j^{(1)}) > t\right) + O(|R_{ij}|) + o(1), \tag{93}$$

in which $Z_i^{(2)}$ and $Z_j^{(2)}$ are two independent random variables (also independent to everything else) following the standard normal distribution. We can further decompose the first term in Equation (93) into four terms as Equation (89) by conditioning on the signs of $Z_i^{(2)}$ and $Z_j^{(2)}$, and repeat the upper bound in Equation (91). This leads to

$$\mathbb{P}(M_i > t, M_j > t) \leq H^2(t) + O(|R_{ij}|) + o(1). \tag{94}$$

Similarly, we can establish the corresponding lower bound. This completes the proof of Lemma A.7.

Corollary A.2. Under Assumption 3.1, as $n, p \to \infty$, for any $t \in \mathbb{R}$, we have

$$\left|\frac{1}{p_0}\sum_{j\in S_0} \mathbb{1}(M_j > t) - H(t)\right| \xrightarrow{p} 0. \tag{95}$$

Proof of Corollary A.2. By Lemma A.7 and Lemma A.5, we have

$$\mathbb{E}\left[\left(\frac{1}{p_0}\sum_{j\in S_0} \mathbb{1}(M_j > t) - H(t)\right)^2\right] \leq \frac{1}{4p_0} + O(||R_{S_0}||_1) + o(1). \tag{96}$$

Under Assumption 3.1, we have $||R_{S_0}||_1 \to 0$ following the arguments in the proof of Lemma A.2 in Zhao et al. (2020). Thus we complete the proof of Corollary A.2 using the Markov inequality.

A.3.2 Proof of Proposition 3.2

For the ease of presentation, we introduce the following notations. For $t \in \mathbb{R}$, denote

$$\widehat{G}_p^0(t) = \frac{1}{p_0}\sum_{j\in S_0} \mathbb{1}(M_j > t), \quad \widehat{G}_p^1(t) = \frac{1}{p_1}\sum_{j\in S^\star} \mathbb{1}(M_j > t), \quad \widehat{V}_p^0(t) = \frac{1}{p_0}\sum_{j\in S_0} \mathbb{1}(M_j < -t). \tag{97}$$

Let $r_p = p_1/p_0$. Denote

$$\text{FDP}_p(t) = \frac{\widehat{G}_p^0(t)}{\widehat{G}_p^0(t) + r_p\widehat{G}_p^1(t)}, \quad \text{FDP}_p^\dagger(t) = \frac{\widehat{V}_p^0(t)}{\widehat{G}_p^0(t) + r_p\widehat{G}_p^1(t)}, \quad \overline{\text{FDP}}_p(t) = \frac{H(t)}{H(t) + r_p\widehat{G}_p^1(t)}. \tag{98}$$

Lemma A.8. We have as $n, p \to \infty$,

$$\begin{aligned} \sup_{t\in\mathbb{R}} \left|\widehat{G}_p^0(t) - H(t)\right| &\xrightarrow{p} 0, \\ \sup_{t\in\mathbb{R}} \left|\widehat{V}_p^0(t) - H(t)\right| &\xrightarrow{p} 0. \end{aligned} \tag{99}$$

Proof of Lemma A.8. We prove the first claim based on an ϵ-net argument. The second claim follows similarly. For any $\epsilon \in (0,1)$, denote $-\infty = \alpha_0^p < \alpha_1^p < \cdots < \alpha_{N_\epsilon}^p = \infty$ in which $N_\epsilon = \lceil 2/\epsilon \rceil$, such that $H(\alpha_{k-1}^p) - H(\alpha_k^p) \le \epsilon/2$ for $k \in [N_\epsilon]$. Such a sequence $\{\alpha_k^p\}$ exists because $H(t)$ is continuous and in the range of $[0,1]$. We have

$$\begin{aligned} \mathbb{P}\left(\sup_{t\in\mathbb{R}} \widehat{G}_p^0(t) - H(t) > \epsilon\right) &\le \mathbb{P}\left(\bigcup_{k=1}^{N_\epsilon} \sup_{t\in[\alpha_{k-1}^p, \alpha_k^p)} \widehat{G}_p^0(t) - H(t) > \epsilon\right) \\ &\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\sup_{t\in[\alpha_{k-1}^p, \alpha_k^p)} \widehat{G}_p^0(t) - H(t) > \epsilon\right). \end{aligned} \tag{100}$$

We note that both $\widehat{G}_p^0(t)$ and $H(t)$ are monotonic decreasing function. Therefore, for any $k \in [N_\epsilon]$, we have

$$\begin{aligned} \sup_{t\in[\alpha_{k-1}^p, \alpha_k^p)} \widehat{G}_p^0(t) - H(t) &\le \widehat{G}_p^0(\alpha_{k-1}^p) - H(\alpha_k^p) \\ &\le \widehat{G}_p^0(\alpha_{k-1}^p) - H(\alpha_{k-1}^p) + \epsilon/2. \end{aligned} \tag{101}$$

By Corollary A.2, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{t\in\mathbb{R}} \widehat{G}_p^0(t) - H(t) > \epsilon\right) &\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\widehat{G}_p^0(\alpha_{k-1}^p) - H(\alpha_{k-1}^p) > \frac{\epsilon}{2}\right) \\ &\le N_\epsilon \max_{k\in[N_\epsilon]} \mathbb{P}\left(\widehat{G}_p^0(\alpha_{k-1}^p) - H(\alpha_{k-1}^p) > \frac{\epsilon}{2}\right) \\ &\longrightarrow 0, \end{aligned} \tag{102}$$

as $n, p \to \infty$. Similarly, we can show that

$$\begin{aligned} \mathbb{P}\left(\inf_{t\in\mathbb{R}} \widehat{G}_p^0(t) - H(t) < -\epsilon\right) &\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\widehat{G}_p^0(\alpha_k^p) - H(\alpha_k^p) < -\frac{\epsilon}{2}\right) \\ &\le N_\epsilon \max_{k\in[N_\epsilon]} \mathbb{P}\left(\widehat{G}_p^0(\alpha_k^p) - H(\alpha_k^p) < -\frac{\epsilon}{2}\right) \\ &\longrightarrow 0. \end{aligned} \tag{103}$$

This concludes the proof of the first claim in Lemma A.8.

Proof of Proposition 3.2. We first show that for any $\epsilon \in (0, q)$, we have

$$\mathbb{P}(\tau_q \le t_{q-\epsilon}) \ge 1 - \epsilon, \tag{104}$$

in which $t_{q-\epsilon} > 0$ satisfying $\mathrm{FDP}^\infty(t_{q-\epsilon}) \le q - \epsilon$. By Lemma A.8, for any fixed $t \in \mathbb{R}$, we have

$$|\mathrm{FDP}_p^\dagger(t) - \mathrm{FDP}_p(t)| \xrightarrow{p} 0. \tag{105}$$

It follows that for any fixed $t \in \mathbb{R}$, we have

$$|\mathrm{FDP}_p^\dagger(t) - \mathrm{FDP}^\infty(t)| \le |\mathrm{FDP}_p^\dagger(t) - \mathrm{FDP}_p(t)| + |\mathrm{FDP}_p(t) - \mathrm{FDP}^\infty(t)| \xrightarrow{p} 0. \tag{106}$$

By the definition of τ_q, i.e., $\tau_q = \inf\{t > 0 : \mathrm{FDP}_p^\dagger(t) \le q\}$, we have

$$\begin{aligned}\mathbb{P}(\tau_q \le t_{q-\epsilon}) &\ge \mathbb{P}(\mathrm{FDP}_p^\dagger(t_{q-\epsilon}) \le q) \\ &\ge \mathbb{P}(|\mathrm{FDP}_p^\dagger(t_{q-\epsilon}) - \mathrm{FDP}^\infty(t_{q-\epsilon})| \le \epsilon) \\ &\ge 1-\epsilon\end{aligned} \tag{107}$$

for p large enough. Conditioning on the event $\tau_q \le t_{q-\epsilon}$, we have

$$\begin{aligned}\limsup_{p\to\infty} \mathbb{E}\left[\mathrm{FDP}_p\left(\tau_q\right)\right] &\le \limsup_{p\to\infty} \mathbb{E}\left[\mathrm{FDP}_p\left(\tau_q\right) \mid \tau_q \le t_{q-\epsilon}\right] \mathbb{P}(\tau_q \le t_{q-\epsilon}) + \epsilon \\ &\le \limsup_{p\to\infty} \mathbb{E}\Big[\Big|\mathrm{FDP}_p\left(\tau_q\right) - \overline{\mathrm{FDP}}_p\left(\tau_q\right)\Big| \;\Big|\; \tau_q \le t_{q-\epsilon}\Big] \mathbb{P}(\tau_q \le t_{q-\epsilon}) \\ &+ \limsup_{p\to\infty} \mathbb{E}\left[\left|\mathrm{FDP}_p^\dagger\left(\tau_q\right) - \overline{\mathrm{FDP}}_p\left(\tau_q\right)\right| \;\middle|\; \tau_q \le t_{q-\epsilon}\right] \mathbb{P}(\tau_q \le t_{q-\epsilon}) \\ &+ \limsup_{p\to\infty} \mathbb{E}\left[\mathrm{FDP}_p^\dagger\left(\tau_q\right) \;\middle|\; \tau_q \le t_{q-\epsilon}\right] \mathbb{P}(\tau_q \le t_{q-\epsilon}) + \epsilon \\ &\le \limsup_{p\to\infty} \mathbb{E}\Big[\sup_{0<t\le t_{q-\epsilon}} \left|\mathrm{FDP}_p(t) - \overline{\mathrm{FDP}}_p(t)\right|\Big] \\ &+ \limsup_{p\to\infty} \mathbb{E}\Big[\sup_{0<t\le t_{q-\epsilon}} \left|\mathrm{FDP}_p^\dagger(t) - \overline{\mathrm{FDP}}_p(t)\right|\Big] \\ &+ \limsup_{p\to\infty} \mathbb{E}\left[\mathrm{FDP}_p^\dagger\left(\tau_q\right)\right] + \epsilon.\end{aligned} \tag{108}$$

The first two terms are 0 based on Lemma A.8 and the dominated convergence theorem. For the third term, we have $\mathrm{FDP}_p^\dagger\left(\tau_q\right) \le q$ almost surely based on the definition of τ_q. This concludes the proof of Proposition 3.2.

A.4 Proof of Proposition 3.3

Similar to Equation (7), we can reparametrize the GLM with respect to the features (X_{-j}, X_j^+, X_j^-) so that the corresponding ture regression coefficients are $\beta_{-j}^\star$, $\beta_j^\star/2$ and $\beta_j^\star/2$, respectively. In addition, the asymptotic signal strength γ and the sampling ratio κ remain the same, whereas the condition variance becomes

$$\mathrm{Var}(X_j^+ \mid X_{-j}, X_j^-) = \mathrm{Var}(X_j^- \mid X_{-j}, X_j^+) = 1/\Theta_{11}^*, \tag{109}$$

in which Θ^* is defined in Proposition 3.3. The proof of Proposition 3.3 thus follows from Lemma A.1 and the proof of Theorem 3.1 in Zhao et al. (2020).

A.5 Proof of Proposition 3.4

The proof of Proposition 3.4 is essentially the same as the proof of Proposition 3.2, once we have the following Lemmas (in particular, Lemma A.13).

Lemma A.9. For the response vector y, we consider fitting two GLMs with respect to the set of features $X = (Z, X_1, \ldots, X_p) \in \mathbb{R}^{n\times(p+1)}$ (full model), and $\widetilde{X} = (X_1, \ldots, X_p) \in \mathbb{R}^{n\times p}$ (reduced model), respectively, in which each row of $\widetilde{X}$ are i.i.d. samples from $N(0, I_p)$, and Z is a null feature following $N(0, I_n)$. Denote the MLEs for the full model and the reduced model as $\widehat{\beta} \in \mathbb{R}^{p+1}$ and $\widetilde{\beta} \in \mathbb{R}^p$, respectively. Denote the difference between the two MLEs as $\Delta = \widehat{\beta}_{2:(p+1)} - \widetilde{\beta}$, then we have

$$||\Delta||_\infty = O_p\big(n^{-1+o(1)}\big).$$

Proof of Lemma A.9. The proof relies heavily on the theoretical results derived in Sur et al. (2019) (see Section 7 therein), and the main technical tool is the leave-one-out analysis. In the following, we sketch the proof of Lemma A.9, and refer the readers to the results in Sur et al. (2019) for complete details.

We first construct a surrogate of $\widehat{\beta}$ as below following Equation (86) in Sur et al. (2019),

$$\widetilde{b} = \begin{bmatrix} 0 \\ \widetilde{\beta} \end{bmatrix} + \widetilde{b}_1 \begin{bmatrix} 1 \\ -\widetilde{G}^{-1} w \end{bmatrix}, \tag{110}$$

in which

$$\widetilde{G} = \frac{1}{n} \widetilde{X}^\intercal D_{\widetilde{\beta}} \widetilde{X}, \qquad w = \frac{1}{n} \widetilde{X}^\intercal D_{\widetilde{\beta}} Z, \tag{111}$$

where $D_{\widetilde{\beta}}$ is a $n \times n$ diagonal matrix with entries $\rho''(\widetilde{x}_i^\intercal \widetilde{\beta})$ for $i \in [n]$, and $\widetilde{b}_1$ is defined as below following Equation (91) in Sur et al. (2019),

$$\widetilde{b}_1 = \frac{Z^\intercal \widetilde{r}}{Z^\intercal D_{\widetilde{\beta}}^{1/2} H D_{\widetilde{\beta}}^{1/2} Z}, \tag{112}$$

where

$$H = I - \frac{1}{n} D_{\widetilde{\beta}}^{1/2} \widetilde{X} \widetilde{G}^{-1} \widetilde{X}^\intercal D_{\widetilde{\beta}}^{1/2} \quad \text{and} \quad \widetilde{r}_i = y_i - \rho'(\widetilde{x}_i^\intercal \widetilde{\beta}), \quad i \in [n]. \tag{113}$$

By Theorem 8 and similar arguments in Section 7.4 in Sur et al. (2019), we have

$$\begin{aligned} ||\Delta||_\infty &\leq ||\widehat{\beta} - \widetilde{b}||_\infty + ||\widetilde{b}_1 \widetilde{G}^{-1} w||_\infty \\ &\lesssim n^{-1+o(1)} + n^{-1/2+o(1)} ||\widetilde{G}^{-1} w||_\infty. \end{aligned} \tag{114}$$

Thus it remains to bound $||\widetilde{G}^{-1} w||_\infty$.

We note that both $\widetilde{X}$ and $D_{\widetilde{\beta}}$ are independent to Z. Therefore, conditioning on $\widetilde{X}$, $\widetilde{G}^{-1} w$ follows a multivariate normal distribution with mean 0, and covariance matrix Σ_Z specified as below,

$$\begin{aligned} \Sigma_Z = \frac{1}{n^2} \widetilde{G}^{-1} \widetilde{X}^\intercal D_{\widetilde{\beta}}^2 \widetilde{X} \widetilde{G}^{-1} &\prec \sup_{i \in [n]} |\rho''(\widetilde{x}_i^\intercal \widetilde{\beta})| \frac{1}{n^2} \widetilde{G}^{-1} \widetilde{X}^\intercal D_{\widetilde{\beta}} \widetilde{X} \widetilde{G}^{-1} \\ &= \frac{1}{n} \sup_{i \in [n]} |\rho''(\widetilde{x}_i^\intercal \widetilde{\beta})| \widetilde{G}^{-1}. \end{aligned} \tag{115}$$

By Lemma 7 in Sur et al. (2019), we have $\sigma_{\min}(\widetilde{G}) > \lambda$ with high probability, in which λ is some positive constant. This implies $\sigma_{\max}(\Sigma_Z) = O_p(1/n)$, which further leads to $||\widetilde{G}^{-1} w||_\infty \lesssim \sqrt{\log n / n}$. Thus we complete the proof of Lemma A.9 via

$$||\Delta||_\infty \lesssim n^{-1+o(1)} + n^{-1/2+o(1)} \sqrt{\log n / n} = O_p(n^{-1+o(1)}). \tag{116}$$

Remark A.1. Lemma A.9 also holds in the case where each row of the design matrix $\widetilde{X}$ independently follows a multivariate normal distribution with a general covariance matrix Σ satisfying Assumption 3.1 (1). In addition, by Lemma A.9, we know that the infinity norm of the absolute difference between the normalized MLEs (defined similarly as Equation (64)) of the full model and the reduced model behave as $\sqrt{n}||\Delta||_\infty = O_p(n^{-1/2+o(1)}) = o(1)$, which is crucial in the proof of Lemma A.13.

Lemma A.10. As $n, p \to \infty$, we have

$$\max_{j \in [p]} |c_j - \tau_j| = o_p(1), \tag{117}$$

in which c_j is defined in Equation (8), and $\tau_j = \mathrm{Var}(X_j | X_{-j})$.

Proof of Lemma A.10. Recall that $c_j = ||P_{-j}^{\perp} X_j||/||P_{-j}^{\perp} Z_j||$. We note that $||P_{-j}^{\perp} Z_j||^2 \sim \chi^2_{n-p+1}$ and $||P_{-j}^{\perp} X_j||^2 \sim \tau_j^2 \chi^2_{n-p+1}$. In addition, $P_{-j}^{\perp} Z_j$ and $P_{-j}^{\perp} X_j$ are independent since Z_j is independent to the design matrix X. The proof is thus completed by the Hoeffiding's inequality and the union bound over $j \in [p]$.

Lemma A.11. For $j \in [p]$, we consider fitting the GLM using the response vector y with respect to two augmented sets of features, $(X_{-j}, X_j + c_j Z_j, X_j - c_j Z_j)$ and $(X_{-j}, X_j + \tau_j Z_j, X_j - \tau_j Z_j)$, in which $\tau_j = \mathrm{Var}(X_j | X_{-j})$, and c_j is defined in Equation (8). Denote the MLEs associated with features $X_j + c_j Z_j, X_j - c_j Z_j$ and $X_j + \tau_j Z_j, X_j - \tau_j Z_j$ as $\widehat{\beta}_j^+, \widehat{\beta}_j^-$ and ${}^*\widehat{\beta}_j^+, {}^*\widehat{\beta}_j^-$, respectively. Further, we denote their normalized versions, defined similarly as Equation (64), as T_j^+, T_j^- and ${}^*\widehat{T}_j^+, {}^*\widehat{T}_j^-$. Then under Assumption 3.1, as $n, p \to \infty$ we have

$$\sup_{j \in [p],\ t_1, t_2 \in \mathbb{R}} \left| \mathbb{P}(T_j^+ > t_1, T_j^- > t_2) - \mathbb{P}({}^*\widehat{T}_j^+ > t_1, {}^*\widehat{T}_j^- > t_2) \right| \longrightarrow 0. \tag{118}$$

Proof of Lemma A.11. We consider fitting the GLM using the response vector y with respect to the augmented set of features (X_{-j}, X_j, Z_j). Denote the MLEs as ${}^\dagger\widehat{\beta}_{-j}, {}^\dagger\widehat{\beta}_j, {}^\dagger\widehat{\vartheta}$. We have the following relationship,

$$\begin{aligned} \widehat{\beta}_j^+ + \widehat{\beta}_j^- &= {}^\dagger\widehat{\beta}_j, \qquad {}^*\widehat{\beta}_j^+ + {}^*\widehat{\beta}_j^- = {}^\dagger\widehat{\beta}_j, \\ \widehat{\beta}_j^+ - \widehat{\beta}_j^- &= {}^\dagger\widehat{\vartheta}/c_j, \quad {}^*\widehat{\beta}_j^+ - {}^*\widehat{\beta}_j^- = {}^\dagger\widehat{\vartheta}/\tau_j. \end{aligned} \tag{119}$$

By Proposition 3.1, $|{}^\dagger\widehat{\vartheta}| = O_p(1/\sqrt{n})$. By Equation (68) and Lemma A.3, A.10, we have

$$\max_{j \in [p]} \left| \widehat{T}_j^+ - {}^*\widehat{T}_j^+ \right| = \max_{j \in [p]} |\widehat{\tau}_j/\sigma_\star| \, |(1/c_j - 1/\tau_j)/2| \, |\sqrt{n}\,{}^\dagger\widehat{\vartheta}| = o_p(1). \tag{120}$$

Similarly, we have $\max_{j \in [p]} |\widehat{T}_j^- - {}^*\widehat{T}_j^-| = o_p(1)$. The proof of Lemma A.11 thus completes based on similar arguments as in the proof of Lemma A.6.

Remark A.2. Without loss of generality, by Lemma A.11, instead of calculating c_j by Equation (8), we assume $c_j = \tau_j$ for $j \in [p]$ henceforth in order to simplify the proofs.

Lemma A.12. Under Assumption 3.1, as $n, p \to \infty$, we have

$$\sup_{t \in \mathbb{R}, j \in S_0} |\mathbb{P}(M_j > t) - H(t)| \to 0. \tag{121}$$

Proof of Lemma A.12. Recall that we fit a GLM with respect to the augmented set of features (X_{-j}, X_j^+, X_j^-), of which the augmented covariance matrix and precision matrix are denoted as Σ_{aug} and Θ_{aug}, respectively. Based on Σ_{aug} and Θ_{aug}, we define V_j^+, V_j^- and $\widetilde{V}_j^+, \widetilde{V}_j^-$ similarly as Equation (63). In particular, V_j^+ and V_j^- are independent since $c_j = \tau_j$. Besides, the normalized MLEs T_j^+, T_j^-, as well as their approximations $\widetilde{T}_j^+, \widetilde{T}_j^-$, are defined similarly as Equation (64).

For $j \in S_0$, we have

$$\begin{aligned} \mathbb{P}(M_j > t) &= \mathbb{P}\Big(\mathrm{sign}\big(T_j^+ T_j^-\big) f\big(T_j^+, T_j^-\big) > t\Big) \\ &= \mathbb{P}\Big(\mathrm{sign}\big(V_j^+ V_j^-\big) f\big(V_j^+, V_j^-\big) > t\Big) + o(1) \\ &= H(t) + o(1). \end{aligned} \tag{122}$$

The second line follows from Corollary A.1, in which the asymptotic vanishing term $o(1)$ is uniform over $j \in S_0$ and $t \in \mathbb{R}$. The proof of Lemma A.12 is thus completed.

Lemma A.13. Under Assumption 3.1, as $n, p \to \infty$, we have

$$\sup_{t\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{p_0}\sum_{j\in S_0}\mathbb{1}(M_j > t)\right) \le \frac{1}{4p_0} + O(||R_{S_0}||_1) + o(1), \tag{123}$$

in which $||R_{S_0}||_1 = \sum_{i,j\in S_0} R_{ij}/p_0^2$.

Proof of lemma A.13. For any $j \in S_0$, let T_j^+, T_j^- be the normalized MLEs, defined similarly as Equation (64), when we fit a GLM with respect to the augmented set of features (X_{-j}, X_j^+, X_j^-). For any different $i, j \in S_0$, let ${}^*T_i^+, {}^*T_i^-$ and ${}^*T_j^+, {}^*T_j^-$ be the normalized MLEs when we fit a GLM with respect to the augmented set of features $(X_{-[i,j]}, X_i^+, X_i^-, X_j^+, X_j^-)$, in which $X_{-[i,j]}$ denotes the design matrix excluding the i-th and j-th columns. Correspondingly, we define ${}^*V_j^+, {}^*V_j^-$ and ${}^*\widetilde{V}_j^+, {}^*\widetilde{V}_j^-$ following Equation (63), and define ${}^*\widetilde{T}_j^+, {}^*\widetilde{T}_j^-$ following Equation (64).

For any different $i, j \in S_0$ and $t > 0$, we have

$$\begin{aligned}
\mathbb{P}(M_i > t, M_j > t) &= \mathbb{P}\Big(\mathrm{sign}\big(T_i^+T_i^-\big)f\big(T_i^+, T_i^-\big) > t, \mathrm{sign}\big(T_j^+T_j^-\big)f\big(T_j^+, T_j^-\big) > t\Big) \\
&= \mathbb{P}\Big(\mathrm{sign}\big({}^*\widetilde{T}_i^{+*}\widetilde{T}_i^-\big)f\big({}^*\widetilde{T}_i^+, {}^*\widetilde{T}_i^-\big) > t\ , \mathrm{sign}\big({}^*\widetilde{T}_j^{+*}\widetilde{T}_j^-\big)f\big({}^*\widetilde{T}_j^+, {}^*\widetilde{T}_j^-\big) > t\Big) + o(1) \\
&= \mathbb{P}\Big(\mathrm{sign}\big({}^*V_i^{+*}V_i^-\big)f\big({}^*V_i^+, {}^*V_i^-\big) > t, \mathrm{sign}\big({}^*V_j^{+*}V_j^-\big)f\big({}^*V_j^+, {}^*V_j^-\big) > t\Big) + o(1),
\end{aligned} \tag{124}$$

in which the second line follows from Lemma A.9, and the last line is based on similar arguments as the proof of Lemma A.6 and Corollary A.1. The rest of proof follows similarly as the proof Lemma A.7, based on a similar decomposition as Equation (89) by conditioning on the signs of ${}^*V_j^+$ and ${}^*V_j^-$. The proof of Lemma A.13 is thus completed.

A.6 Proof of Proposition 4.1

For the design matrix X, we denote X_j as its j-th column and x_k as its k-th row, in which $j \in [p]$ and $k \in [n]$. We introduce the normalized versions of Θ and Λ defined as below,

$$\Theta_{ij}^0 = \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}, \quad \Lambda_{ij}^0 = \frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}}. \tag{125}$$

A.6.1 Technical lemmas

Lemma A.14. Under Assumption 4.1, for any $\epsilon > 0$, there exist constants $c_1, c_2, c_3, c_4 > 0$ such that for large enough n,

$$\mathbb{P}\left(\max_{i,j\in[p]}|\Lambda_{ij} - \Theta_{ij}| \le \frac{1}{\sqrt{\log p}}\right) \ge 1 - c_1p^2\exp\left(-c_2\frac{n}{\log p}\right) - c_3p^2\exp\left(-c_4n\right) - \epsilon. \tag{126}$$

Proof of Lemma A.14. The proof follows similarly as the proof of Lemma 7.2 in Javanmard and Montanari (2013). For any $\epsilon > 0$, by Proposition 5.1 in Javanmard and Montanari (2013) (also see Van de Geer et al. (2014)), there exists an $n_\epsilon \in \mathbb{N}_+$ such that for $n \ge n_\epsilon$, we have

$$\mathbb{P}\left(||\widehat{\Theta} - \Theta||_\infty \le \frac{1}{\sqrt{\log p}}\right) \ge 1 - \epsilon. \tag{127}$$

In the following, we condition on this high probability event. Denote $v = \Theta^\top e_i, u = \Theta^\top e_j, \delta = (\Theta - \hat{\Theta})^\top e_i, \eta = (\Theta - \hat{\Theta})^\top e_j$. We have the following decomposition,

$$\begin{aligned}
\Lambda_{ij} - \Theta_{ij} &= (v - \delta)^\top\hat{\Sigma}(u - \eta) - \Theta_{ij} \\
&= \big(v^\top\hat{\Sigma}u - \Theta_{ij}\big) - v^\top\hat{\Sigma}\eta - \delta^\top\hat{\Sigma}u + \delta^\top\hat{\Sigma}\eta.
\end{aligned} \tag{128}$$

We proceed to bound each term. For the term $v^\top \hat{\Sigma} u - \Theta_{ij}$, since $\mathbb{E}[v^\top \hat{\Sigma} u] = v^\top \Sigma u = \Theta_{ij}$, it follows that

$$\begin{aligned} v^\top \hat{\Sigma} u - \Theta_{ij} &= v^\top \hat{\Sigma} u - \mathbb{E}[v^\top \hat{\Sigma} u] \\ &= \frac{1}{n} \sum_{k=1}^{n} e_i^\top \Theta \left(x_k x_k^\top - \mathbb{E}[x_k x_k^\top] \right) \Theta^\top e_j \end{aligned} \tag{129}$$

Denote $\xi_k = e_i^\top \Theta \left(x_k x_k^\top - \mathbb{E}[x_k x_k^\top] \right) \Theta^\top e_j$ for $k \in [n]$. We note that ξ_k's are independent, and have sub-exponential tails because

$$||\xi_k||_{\psi_1} \leq 2||(e_i^\top \Theta x_k)^2||_{\psi_1} \leq 4||e_i^\top \Theta x_k||_{\psi_2}^2 \leq K \tag{130}$$

for some constant $K > 0$, where the first inequality follows from Vershynin (2010) (Remark 5.18), the second inequality follows from Lemma C.1 in Javanmard and Montanari (2013), and the last inequality follows from Assumption 4.1 (2). Without loss of generality, we assume $K \geq 1$. Using a Bernstein-type inequality (Proposition 5.26 in Vershynin (2010)), we have

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^{n} \xi_k \right| \geq t \right) \leq 2 \exp\left(-cn \min\left(\frac{t^2}{K^2}, \frac{t}{K} \right) \right). \tag{131}$$

Plugging in $t = 1/\sqrt{\log p}$ and employing the union bound, we obtain

$$\mathbb{P}\left(\max_{i,j \in [p]} \left| v^\top \hat{\Sigma} u - \Theta_{ij} \right| \geq \frac{1}{\sqrt{\log p}} \right) \leq 2p^2 \exp\left(-c_2 \frac{n}{\log p} \right) \tag{132}$$

for some constant $c_2 > 0$.

Next we consider the term $\delta^\top \hat{\Sigma} \eta$. Since $\hat{\Sigma} \succeq 0$, it is sufficient for us to bound both $\delta^\top \hat{\Sigma} \delta$ and $\eta^\top \hat{\Sigma} \eta$ by the Cauchy-Schwartz inequality. Notice that

$$\delta^\top \Sigma \delta = \sum_{i,j \in [p]} \hat{\Sigma}_{ij} \delta_i \delta_j \leq |\hat{\Sigma}|_\infty ||\delta||_1^2 \leq |\hat{\Sigma}|_\infty ||\Theta - \hat{\Theta}||_\infty^2 \leq \frac{|\hat{\Sigma}|_\infty}{\log p}. \tag{133}$$

We proceed to bound the tail probability $\mathbb{P}(|\hat{\Sigma}|_\infty \geq 2)$. First, using the union bound, we have

$$\mathbb{P}(|\hat{\Sigma}|_\infty \geq 2) \leq \sum_{i,j \in [p]} \mathbb{P}(|\hat{\Sigma}_{ij}| \geq 2) \leq \sum_{i,j \in [p]} \mathbb{P}(|\hat{\Sigma}_{ij} - \mathbb{E}\hat{\Sigma}_{ij}| \geq 1). \tag{134}$$

Denote $\xi_k = e_i^\top x_k x_k^\top e_j - \Sigma_{ij}$ for $k \in [n]$. We have

$$||\xi_k||_{\psi_1} \leq 2||(e_i^\top x_k)^2||_{\psi_1} \leq 4||e_i^\top x_k||_{\psi_2}^2 \leq 4||e_i^\top \Theta^{-1/2} \Theta^{1/2} x_k||_{\psi_2}^2 \leq K \tag{135}$$

for some constant $K > 0$. Thus ξ_ks are independent sub-exponential random variables. By the Bernstein inequality, we have

$$\mathbb{P}(|\hat{\Sigma}|_\infty \geq 2) \leq 2p^2 \exp(-c_4 n) \tag{136}$$

for some constant $c_4 > 0$. We remark that although $\delta = (\Theta - \hat{\Theta})^\top e_i$ is implicitly associated with the index i, the upper bound $|\hat{\Sigma}|_\infty ||\Theta - \hat{\Theta}||_\infty^2$ in Equation (133) is irrelevant to the index i. Therefore, we have shown that

$$\mathbb{P}\left(\max_{i,j \in [p]} |\delta^\top \hat{\Sigma} \eta| \geq \frac{2}{\log p} \right) \leq 4p^2 \exp(-c_4 n) + 2\epsilon, \tag{137}$$

where the maximum is taken with respect to the implicit index i, j associated with δ and η.

We now consider the term $v^\top \hat{\Sigma}\eta$. We have

$$\begin{aligned}\max_{i,j\in[p]} |v^\top \hat{\Sigma}\eta| &\leq \max_{i,j\in[p]} [v^\top \hat{\Sigma} v]^{1/2} [\eta^\top \hat{\Sigma}\eta]^{1/2} \\ &\leq \max_{i,j\in[p]} [|v^\top \hat{\Sigma} v - \Theta_{ii}| + |\Theta_{ii}|]^{1/2} [\eta^\top \hat{\Sigma}\eta]^{1/2} \\ &\leq [\max_{i\in[p]} |v^\top \hat{\Sigma} v - \Theta_{ii}| + \max_{i\in[p]} |\Theta_{ii}|]^{1/2} \max_{j\in[p]} [\eta^\top \hat{\Sigma}\eta]^{1/2}.\end{aligned} \tag{138}$$

Under Assumption 4.1 (2), $\max_{i\in[p]} |\Theta_{ii}|$ is upper bounded. Combining the inequalities in Equation (132) and Equation (137), we have

$$\mathbb{P}\left(\max_{i,j\in[p]} |v^\top \hat{\Sigma}\eta| \geq \frac{c}{\sqrt{\log p}}\right) \leq 2p^2 \exp\left(-c_2 \frac{n}{\log p}\right) + 2p^2 \exp(-c_4 n) + \epsilon. \tag{139}$$

The decomposition in Equation (128), along with the inequalities in Equation (132), Equation (137) and Equation (139) together imply the claim in Lemma A.14.

Corollary A.3. The high probability bound in Lemma A.14 also applies to the normalized versions Λ^0 and Θ^0. That is, under Assumption 4.1, for any $\epsilon > 0$, there exist constants $c_1, c_2, c_3, c_4 > 0$ such that for large enough n,

$$\mathbb{P}\left(\max_{i,j\in[p]} \left|\Lambda^0_{ij} - \Theta^0_{ij}\right| \leq \frac{1}{\sqrt{\log p}}\right) \geq 1 - c_1 p^2 \exp\left(-c_2 \frac{n}{\log p}\right) - c_3 p^2 \exp\left(-c_4 n\right) - \epsilon. \tag{140}$$

Proof of Corollary A.3. We show that $\max_{i,j\in[p]} \left|\Lambda^0_{ij} - \Theta^0_{ij}\right| \leq c/\sqrt{\log p}$ for some constant $c > 0$, conditioning on the high probability event $\left\{\max_{i,j\in[p]} |\Lambda_{ij} - \Theta_{ij}| \leq 1/\sqrt{\log p}\right\}$. We have

$$\begin{aligned}\max_{i,j\in[p]} \left|\Lambda^0_{ij} - \Theta^0_{ij}\right| &= \max_{i,j\in[p]} \left|\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} - \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}\right| \\ &\leq \max_{i,j\in[p]} \left|\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} - \frac{\Lambda_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}\right| + \max_{i,j\in[p]} \left|\frac{\Lambda_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} - \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}\right|.\end{aligned} \tag{141}$$

Consider the first term. By Assumption 4.1 (2), for large enough p, we have

$$\frac{|\Lambda_{ij}|}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} \leq \frac{|\Lambda_{ij} - \Theta_{ij}| + \Theta_{ij}}{\sqrt{[\Theta_{ii} - |\Lambda_{ii} - \Theta_{ii}|][\Theta_{jj} - |\Lambda_{jj} - \Theta_{jj}|]}} \leq K \tag{142}$$

for some constant $K > 0$. It follows that

$$\begin{aligned}\left|\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} - \frac{\Lambda_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}\right| &= \frac{|\Lambda_{ij}|}{\sqrt{\Lambda_{ii}\Lambda_{jj}}} \left|\left(\frac{\Lambda_{ii}\Lambda_{jj}}{\Theta_{ii}\Theta_{jj}}\right)^{1/2} - 1\right| \leq K \left|\left(\frac{\Lambda_{ii}\Lambda_{jj}}{\Theta_{ii}\Theta_{jj}}\right)^{1/2} - 1\right| \\ &\leq K \left|\left(1 + \frac{1}{\sqrt{\log p}\Theta_{ii}}\right)^{1/2} \left(1 + \frac{1}{\sqrt{\log p}\Theta_{jj}}\right)^{1/2} - 1\right| \\ &\leq K \left|1 + \frac{2C}{\sqrt{\log p}} + \frac{C^2}{\log p} - 1\right| \leq \frac{K'}{\sqrt{\log p}}\end{aligned} \tag{143}$$

for some constant $K' > 0$. In the last to second inequality above, we use an elementary inequality $\sqrt{1+x} \leq 1 + x$ for $x > 0$.

For the second term, by Assumption 4.1 (2), we have

$$\left|\frac{\Lambda_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} - \frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}}\right| \leq C|\Lambda_{ij} - \Theta_{ij}| \leq \frac{C}{\sqrt{\log p}}. \tag{144}$$

We note that both the upper bounds in Equation (143) and Equation (144) do not depend on the indexes i, j. Therefore, we have shown that $\max_{i,j\in[p]} \left|\Lambda^0_{ij} - \Theta^0_{ij}\right| \leq (K' + C)/\sqrt{\log p}$. This implies the claim in Corollary A.3.

Lemma A.15. Under Assumption 4.1, as $n, p \to \infty$, we have

$$\sup_{t\in\mathbb{R},\ j\in S_0} |\mathbb{P}(T_j > t) - Q(t)| \longrightarrow 0. \tag{145}$$

Proof of Lemma A.15. Recall that

$$T_j = \sqrt{n}\widehat{\beta}^d_j/\Lambda^{1/2}_{jj} \quad \text{and} \quad \widetilde{Z}_j := Z_j/\Lambda^{1/2}_{jj} \Big| X \sim N(0,1). \tag{146}$$

Without changing the selection result obtained via Algorithm 1, we multiply the normalized debiased Lasso estimator T_j defined in Equation (23) by a constant factor $\sqrt{n}$ so that it follows the standard normal distribution asymptotically.

By Lemma A.14, for large enough n, we have $\min_{j\in[p]} \Lambda_{jj} \geq c$ for some constant $c > 0$ with high probability. Henceforth we condition on this high probability event. Recall the decomposition in Equation (20), thus for $j \in S_0$, i.e., $\beta^\star_j = 0$, we have $T_j = \widetilde{Z}_j + \Delta_j/\Lambda^{1/2}_{jj}$. Let $\epsilon = p_1 \log p/\sqrt{n} \to 0$ by Assumption 4.1 (1). For the bias term Δ, we have

$$\mathbb{P}\left(||\Delta||_\infty \geq \epsilon\sqrt{c}\right) \leq \epsilon$$

by Theorem 2.3 in Javanmard and Montanari (2013). It follows that for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\left(T_j > t\right) - Q(t) &= \mathbb{P}\left(\widetilde{Z}_j > t - \Delta_j/\Lambda^{1/2}_{jj}\right) - Q(t) \\
&\leq \mathbb{P}(\widetilde{Z}_j > t - \epsilon) + \mathbb{P}\left(||\Delta||_\infty \geq \epsilon\sqrt{c}\right) - Q(t) \\
&\leq Q(t-\epsilon) - Q(t) + \epsilon \\
&\leq c_1\epsilon
\end{aligned} \tag{147}$$

for some constant $c_1 > 0$, in which the last inequality follows from the simple fact that for any $t \in \mathbb{R},\ \epsilon > 0$, we have

$$Q(t-\epsilon) - Q(t) = \int_{t-\epsilon}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \frac{\epsilon}{\sqrt{2\pi}}. \tag{148}$$

Similarly, we have

$$\begin{aligned}
\mathbb{P}\left(T_j > t\right) - Q(t) &\geq \mathbb{P}(\widetilde{Z}_j > t + \epsilon)\mathbb{P}(||\Delta||_\infty < \epsilon\sqrt{c}) - Q(t) \\
&\geq Q(t+\epsilon)(1-\epsilon) - Q(t) \\
&\geq -c_2\epsilon.
\end{aligned} \tag{149}$$

for some constant $c_2 > 0$. Since the bounds in Equation (147) and Equation (149) are irrelevant to t and the index j, the claim in Lemma A.15 follows.

Lemma A.16. Under Assumption 4.1, as $n, p \to \infty$, we have

$$\sup_{t\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{p_0}\sum_{j\in S_0} \mathbb{1}(M_j > t)\right) \longrightarrow 0. \tag{150}$$

Proof of Lemma A.16. Denote the correlated set as $\Gamma = \{(i,j) : i,j \in S_0,\ \Theta_{ij} \neq 0\}$, and the uncorrelated set as $\Gamma^c = \{(i,j) : i,j \in S_0,\ \Theta_{ij} = 0\}$. We have the following decomposition.

$$\begin{aligned}\sup_{t\in\mathbb{R}} \mathrm{Var}\left(\frac{1}{p_0}\sum_{j\in S_0}\mathbb{1}(M_j > t)\right) &\leq \frac{1}{p_0^2}\sup_{t\in\mathbb{R}}\sum_{(i,j)\in\Gamma}\mathrm{Cov}(\mathbb{1}(M_i > t), \mathbb{1}(M_j > t)) \\ &\quad + \frac{1}{p_0^2}\sup_{t\in\mathbb{R}}\sum_{(i,j)\in\Gamma^c}\mathrm{Cov}(\mathbb{1}(M_i > t), \mathbb{1}(M_j > t)) \\ &\leq \frac{|\Gamma|}{p_0^2} + \sup_{t\in\mathbb{R},(i,j)\in\Gamma^c}\left|\mathbb{P}(M_i > t, M_j > t) - H^2(t)\right| \\ &\quad + \sup_{t\in\mathbb{R},(i,j)\in\Gamma^c}\left|\mathbb{P}(M_i > t)\mathbb{P}(M_j > t) - H^2(t)\right|.\end{aligned} \tag{151}$$

By Assumption 4.1 (1), we have $|\Gamma|/p_0^2 \leq p_0\sqrt{n}/(p_0^2\log p) \to 0$. Further, by Lemma A.5, we have

$$\sup_{t\in\mathbb{R},i,j\in S_0}\left|\mathbb{P}(M_i > t)\mathbb{P}(M_j > t) - H^2(t)\right| \to 0. \tag{152}$$

Therefore, it is sufficient for us to show that

$$\sup_{t\in\mathbb{R},(i,j)\in\Gamma^c}\left|\mathbb{P}(M_i > t, M_j > t) - H^2(t)\right| \to 0, \quad \text{as } n,p\to\infty. \tag{153}$$

In the following, we condition on the high probability event $E_1 \cap E_2 \cap E_3$, in which

$$\begin{aligned}E_1 &= \Big\{\min_{j\in[p]}\Lambda^{(1)}_{jj} \geq c\Big\} \cap \Big\{\min_{j\in[p]}\Lambda^{(2)}_{jj} \geq c\Big\}, \\ E_2 &= \Big\{\max_{i,j\in[p]}|\Lambda^{0,(1)}_{ij} - \Theta^0_{ij}| \leq \frac{1}{\sqrt{\log p}}\Big\} \cap \Big\{\max_{i,j\in[p]}|\Lambda^{0,(2)}_{ij} - \Theta^0_{ij}| \leq \frac{1}{\sqrt{\log p}}\Big\}, \\ E_3 &= \Big\{||\Delta^{(1)}||_\infty \leq \epsilon\sqrt{c}\Big\} \cap \Big\{||\Delta^{(2)}||_\infty \leq \epsilon\sqrt{c}\Big\}.\end{aligned} \tag{154}$$

The superscripts $(1),(2)$ reflects the data splitting step, and the corresponding random variables are defined on each part of the data. $c > 0$ is some suitable constant (see Lemma A.14 and Assumption 4.1 (2)), and $\epsilon = p_1\log p/\sqrt{n} \to 0$ by Assumption 4.1 (1).

We consider the same decomposition as in Equation (89), and proceed to upper bound I_1. Let $\rho = 1/\sqrt{\log p}$. Notice that for $(i,j) \in \Gamma^c$, $\Theta^0_{ij} = 0$, thus we have

$$\max_{(i,j)\in\Gamma^c}\left|\Lambda^{0,(1)}_{i,j}\right| \leq \rho \quad \text{and} \quad \max_{(i,j)\in\Gamma^c}\left|\Lambda^{0,(2)}_{i,j}\right| \leq \rho$$

once we condition on the high probability event E_2. Let $\widetilde{Z}^{(1)}$ and $\widetilde{Z}^{(2)}$ be the normalized version (e.g. variance 1) of $Z^{(1)}$ and $Z^{(2)}$ (see Equation (20)), respectively, thus $\Lambda^{0,(1)}$ and $\Lambda^{0,(2)}$ are essentially the corresponding covariance matrices. We have

$$\begin{aligned}I_1 &\leq \mathbb{P}\left(\widetilde{Z}^{(2)}_i > I_t(T^{(1)}_i) - \epsilon, \widetilde{Z}^{(2)}_j > I_t(T^{(1)}_j) - \epsilon, T^{(1)}_i > 0, T^{(1)}_j > 0\right) \\ &= \mathbb{E}\left[\mathbb{P}\left(\widetilde{Z}^{(2)}_i > I_t(x) - \epsilon, \widetilde{Z}^{(2)}_j > I_t(y) - \epsilon\right) \mid T^{(1)}_i = x > 0,\ T^{(1)}_j = y > 0\right] \\ &\leq c_1\rho + \mathbb{E}\left[Q\big(I_t(x) - \epsilon\big)Q\big(I_t(y) - \epsilon\big) \mid T^{(1)}_i = x > 0,\ T^{(1)}_j = y > 0\right] \\ &\leq c_1\rho + c_2\epsilon + \mathbb{E}\left[Q\big(I_t(x)\big)Q\big(I_t(y)\big) \mid T^{(1)}_i = x > 0,\ T^{(1)}_j = y > 0\right]\end{aligned} \tag{155}$$

for some constants $c_1, c_2 > 0$, in which the third line is based on the Mehler's identity (see Equation (90)) and Lemma 1 in Azriel and Schwartzman (2015), and the last line is based on the inequality

in Equation (148). Similarly, we can upper bound I_2, I_3 and I_4. Combining the four upper bounds together, we obtain an upper bound on $\mathbb{P}(M_i > t, M_j > t)$ specified as below,

$$\mathbb{P}\left(\text{sign}(W_i^{(2)}T_i^{(1)})f(W_i^{(2)}, T_i^{(1)}) > t, \text{sign}(W_j^{(2)}T_j^{(1)})f(W_j^{(2)}, T_j^{(1)}) > t\right) + 4c_1\rho + 4c_2\epsilon, \tag{156}$$

in which $W_i^{(2)}$ and $W_j^{(2)}$ are two independent random variables (also independent to everything else) following the standard normal distribution. We can further decompose the first term in Equation (156) into four terms as Equation (89) by conditioning on the signs of $W_i^{(2)}$ and $W_j^{(2)}$, and repeat the upper bound in Equation (155). This leads to

$$\mathbb{P}(M_i > t, M_j > t) \leq H^2(t) + 8c_1\rho + 8c_2\epsilon. \tag{157}$$

Similarly, we can establish the corresponding lower bound. Thus we complete the proof of Lemma A.16.

A.6.2 Proof of Proposition 4.1

In addition to the notations introduced in Section A.3.2, we denote

$$G_p^0(t) = \frac{1}{p_0}\sum_{j\in S_0}\mathbb{P}(M_j > t), \quad V_p^0(t) = \frac{1}{p_0}\sum_{j\in S_0}\mathbb{P}(M_j < -t). \tag{158}$$

The proof of Proposition 4.1 is essentially the same as the proof of Proposition 3.2, with the help of the following Lemma A.17.

Lemma A.17. Under Assumption 4.1, as $n, p \to \infty$, we have in probability,

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\widehat{G}_p^0(t) - H(t)\right| &\longrightarrow 0,\\ \sup_{t\in\mathbb{R}}\left|\widehat{V}_p^0(t) - H(t)\right| &\longrightarrow 0.\end{aligned} \tag{159}$$

Proof of Lemma A.17. We prove the first claim. The second claim follows similarly. Notice that $H(t)$ is symmetric about 0. By Lemma A.5, it is sufficient for us to show that

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\widehat{G}_p^0(t) - G_p^0(t)\right| &\longrightarrow 0,\\ \sup_{t\in\mathbb{R}}\left|\widehat{V}_p^0(t) - V_p^0(t)\right| &\longrightarrow 0.\end{aligned} \tag{160}$$

The proof follows similarly as the proof of Lemma A.8 using an ϵ-net argument, except that we use the Chebyshev's inequality and Lemma A.16 in Equation (102).

A.7 Proof of Proposition 4.2 and 4.3

Without loss of generality, we assume that the Lipschitz constant of $\ddot{\rho}(v)$ is 1 (see Assumption 4.2 (3)). For $j \in [p]$, denote $\eta_j = X_{\beta^\star,j} - X_{\beta^\star,-j}\gamma_j$, and denote $\tau_j^2 = \mathbb{E}[\eta_j^\top\eta_j/n]$ as the conditional variance $\text{Var}(X_{\beta^\star,j}|X_{\beta^\star,-j})$. Denote $\widehat{\Sigma} = P_n\dot{\ell}_{\widehat{\beta}}\dot{\ell}_{\widehat{\beta}}^\top$ as the sample version of Σ.

A.7.1 Technical lemmas

Lemma A.18. Under Assumption 4.2, for any $\epsilon > 0$, there exists a constant $C_\epsilon > 0$, such that $\mathbb{P}(\cap_{i=1}^4 E_i^\epsilon) \geq 1 - \epsilon$, in which the events $E_1^\epsilon, E_2^\epsilon, E_3^\epsilon, E_4^\epsilon$ are defined as below,

$$\begin{aligned} E_1^\epsilon &= \left\{ \frac{1}{n} ||X(\widehat{\beta} - \beta^\star)||_2^2 \leq C_\epsilon p_1 \log p/n \right\}, \\ E_2^\epsilon &= \left\{ ||\widehat{\beta} - \beta^\star||_1 \leq C_\epsilon p_1 \sqrt{\log p/n} \right\}, \\ E_3^\epsilon &= \left\{ \max_{j \in [p]} ||\widehat{\gamma}_j - \gamma_j||_1 \leq C_\epsilon s \sqrt{\log p/n} \right\}, \\ E_4^\epsilon &= \left\{ \max_{j \in [p]} ||\widehat{\gamma}_j - \gamma_j||_2 \leq C_\epsilon \sqrt{s \log p/n} \right\}. \end{aligned} \tag{161}$$

Proof of Lemma A.18. Lemma A.18 follows from standard arguments in Bickel et al. (2009), Raskutti et al. (2010), Bühlmann and Van de Geer (2011), Van de Geer et al. (2014).

Lemma A.19. Under Assumption 4.2, we have

$$\begin{aligned} &\max_{j \in [p]} \left| \widehat{\tau}_j^2 - \tau_j^2 \right| = O_p\big(\sqrt{s \log p/n}\big), \\ &\max_{j \in [p]} \left| 1/\widehat{\tau}_j^2 - 1/\tau_j^2 \right| = O_p\big(\sqrt{s \log p/n}\big). \end{aligned} \tag{162}$$

Proof of Lemma A.19. The proof follows similarly as the proof of Theorem 3.2 in Van de Geer et al. (2014), combined with some enriched arguments that we detail below. For any $\epsilon > 0$, we condition on the high probability event $\cap_{i=1}^4 E_i^\epsilon$. Since $X_{\widehat{\beta}} = W_{\widehat{\beta}} W_{\beta^\star}^{-1} X_{\beta^\star}$, we have the following decomposition,

$$\widehat{\tau}_j^2 - \tau_j^2 = \frac{1}{n} X_{\widehat{\beta},j}^\top \left(X_{\widehat{\beta},j} - X_{\widehat{\beta},-j} \widehat{\gamma}_j \right) - \tau_j^2 := I_1 + I_2, \tag{163}$$

in which

$$\begin{aligned} I_1 &= \frac{1}{n} X_{\beta^\star,j}^\top \left(X_{\beta^\star,j} - X_{\beta^\star,-j} \widehat{\gamma}_j \right) - \tau_j^2, \\ I_2 &= \frac{1}{n} X_{\beta^\star,j}^\top \left(W_{\widehat{\beta}}^2 W_{\beta^\star}^{-2} - I \right) \left(X_{\beta^\star,j} - X_{\beta^\star,-j} \widehat{\gamma}_j \right). \end{aligned} \tag{164}$$

We proceed to upper bound I_1 and I_2.

We further decompose I_1 into four terms, $I_1 = I_{11} + I_{12} + I_{13} + I_{14}$, in which

$$\begin{aligned} I_{11} &= \frac{1}{n} \eta_i^\top \eta_j - \tau_j^2, \\ I_{12} &= \frac{1}{n} \eta_j^\top X_{\beta^\star,-j} (\gamma_j - \widehat{\gamma}_j), \\ I_{13} &= \frac{1}{n} \gamma_j^\top X_{\widehat{\beta},-j}^\top \left(X_{\widehat{\beta},j} - X_{\widehat{\beta},-j} \widehat{\gamma}_j \right), \\ I_{14} &= \frac{1}{n} \gamma_j^\top X_{\beta^\star,-j}^\top \left(I - W_{\widehat{\beta}}^2 W_{\beta^\star}^{-2} \right) \left(X_{\beta^\star,j} - X_{\beta^\star,-j} \widehat{\gamma}_j \right). \end{aligned} \tag{165}$$

For I_{11}, by Assumption 4.2 (2), we have

$$||\eta_j||_\infty \leq ||X_{\beta^\star,j}||_\infty + ||X_{\beta^\star,-j} \gamma_j||_\infty \leq 2C_1. \tag{166}$$

Since $\mathbb{E}[\eta_j^\top \eta_j / n] = \tau_j^2$, we have

$$\max_{j \in [p]} I_{11} = O_p(\sqrt{\log p/n}), \tag{167}$$

based on the union bound and the Hoeffding's inequality.

For I_{12}, since $\mathbb{E}[\eta_j^\top X_{\beta^\star,-j}] = 0$, we have

$$\max_{j\in[p]} ||\eta_j^\top X_{\beta^\star,-j}/n||_\infty = O_p(\sqrt{\log p/n}), \tag{168}$$

by the union bound and the Hoeffding's inequality. It follows that

$$\begin{aligned}\max_{j\in[p]} I_{12} &\leq \max_{j\in[p]} ||\eta_j^\top X_{\widehat{\beta},-j}/n||_\infty \max_{j\in[p]} ||\widehat{\gamma}_j - \gamma_j||_1 \\ &= O_p(s\log p/n).\end{aligned} \tag{169}$$

For I_{13}, since γ_j is $s-$sparse, by Assumption 4.2 (2), we have

$$\begin{aligned}\max_{j\in[p]} ||\gamma_j||_1 &\leq \sqrt{s}\max_{j\in[p]} ||\gamma_j||_2 \\ &= \sqrt{s}\max_{j\in[p]} [\Sigma_{j,-j}\Sigma_{-j,-j}^{-2}\Sigma_{-j,j}]^{1/2} \\ &\leq \sqrt{s}C_1^2/C_2.\end{aligned} \tag{170}$$

In addition, by the KKT condition of the j-th nodewise Lasso regression, we have

$$\max_{j\in[p]} ||X_{\widehat{\beta},-j}^\top \left(X_{\widehat{\beta},j} - X_{\widehat{\beta},-j}\widehat{\gamma}_j\right)/n||_\infty \leq \lambda_j. \tag{171}$$

It follows that

$$\begin{aligned}\max_{j\in[p]} I_{13} &\leq \max_{j\in[p]} ||\gamma_j||_1 \max_{j\in[p]} ||X_{\widehat{\beta},-j}^\top \left(X_{\widehat{\beta},j} - X_{\widehat{\beta},-j}\widehat{\gamma}_j\right)/n||_\infty \\ &= O_p(\sqrt{s\log p/n}).\end{aligned} \tag{172}$$

For I_{14}, we first notice that

$$\begin{aligned}||X_{\beta^\star,j} - X_{\beta^\star,-j}\widehat{\gamma}_j||_\infty &\leq ||\eta_j||_\infty + ||X_{\beta^\star,-j}(\gamma_j - \widehat{\gamma}_j)||_\infty \\ &\leq 2C_1 + C_1||\gamma_j - \widehat{\gamma}_j||_1 \\ &\leq 3C_1\end{aligned} \tag{173}$$

by Assumption 4.2 (2) and Equation (166). By Assumption 4.2 (3), it follows that

$$\begin{aligned}\max_{j\in[p]} I_{14} &\leq \frac{3C_1^2}{n}\sum_{i=1}^n \left|\frac{\ddot{\rho}(x_i^\intercal\widehat{\beta}) - \ddot{\rho}(x_i^\intercal\beta^\star)}{\ddot{\rho}(x_i^\intercal\beta^\star)}\right| \\ &= O_p(||X(\widehat{\beta} - \beta^\star)||_2/\sqrt{n}) \\ &= O_p(\sqrt{p_1\log p/n}),\end{aligned} \tag{174}$$

in which the second inequality follows from Cauchy-Schwarz inequality. Combining the upper bound on $\max_{j\in[p]} I_{11}$, $\max_{j\in[p]} I_{12}$, $\max_{j\in[p]} I_{13}$ and $\max_{j\in[p]} I_{14}$, we show that

$$\max_{j\in[p]} I_1 = O_p(\sqrt{s\log p/n}). \tag{175}$$

Finally, $\max_{j\in[p]} I_2$ can be upper bounded similarly as $\max_{j\in[p]} I_{14}$. This completes the proof of the first claim in Lemma A.19. The second claim follows by noticing that $1/\tau_j^2$ is uniformly upper bounded because

$$1/\tau_j^2 = \Theta_{j,j} \leq \sigma_{\max}(\Theta) = 1/\sigma_{\min}(\Sigma) \leq C_2. \tag{176}$$

This completes the proof of Lemma A.19

Lemma A.20. Under Assumption 4.2, we have

$$\begin{aligned}
&\max_{j\in[p]} ||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_1 = O_p(s\sqrt{\log p/n}),\\
&\max_{j\in[p]} ||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_2 = O_p(\sqrt{s\log p/n}),\\
&\max_{j\in[p]} |\widehat{\Theta}_{j,\cdot}\Sigma\widehat{\Theta}_{j,\cdot}^\top - \Theta_{j,j}| = O_p(\sqrt{s\log p/n}).
\end{aligned} \tag{177}$$

Proof of Lemma A.20. By Lemma A.18 and A.19, for any $\epsilon > 0$, there exists a constant $C_\epsilon > 0$, such that $\mathbb{P}(\cap_{i=1}^6 E_i^\epsilon) \geq 1-\epsilon$, in which E_i^ϵ for $i \in [4]$ are specified in Lemma A.18, and

$$\begin{aligned}
E_5^\epsilon &= \left\{\max_{j\in[p]} \left|\widehat{\tau}_j^2 - \tau_j^2\right| \leq C_\epsilon\sqrt{s\log p/n}\right\},\\
E_6^\epsilon &= \left\{\max_{j\in[p]} \left|1/\widehat{\tau}_j^2 - 1/\tau_j^2\right| \leq C_\epsilon\sqrt{s\log p/n}\right\}.
\end{aligned} \tag{178}$$

In the following, we condition on this high probability event $\cap_{i=1}^6 E_i^\epsilon$. By Assumption 4.2 (1), we assume that n is large enough so that $C_\epsilon\sqrt{s\log p/n} \leq C_2$.

We first note that by Equation (176),

$$1/\widehat{\tau}_j^2 \leq |1/\widehat{\tau}_j^2 - 1/\tau_j^2| + 1/\tau_j^2 \leq 2C_2. \tag{179}$$

Thus, by Equation (170), we have

$$\begin{aligned}
\max_{j\in[p]} ||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_1 &= \max_{j\in[p]} ||\widehat{C}_{j,\cdot}/\widehat{\tau}_j^2 - C_{j,\cdot}/\tau_j^2||_1\\
&\leq \max_{j\in[p]} ||\widehat{\gamma}_j - \gamma_j||_1 \max_{j\in[p]} 1/\widehat{\tau}_j^2 + \max_{j\in[p]} ||\gamma_j||_1 \max_{j\in[p]} \left|1/\widehat{\tau}_j^2 - 1/\tau_j^2\right|\\
&\leq 2C_2C_\epsilon s\sqrt{\log p/n} + C_1^2/C_2C_\epsilon\sqrt{s}\sqrt{s\log p/n}\\
&= O_p(s\sqrt{\log p/n}).
\end{aligned} \tag{180}$$

Similarly, we have

$$\begin{aligned}
\max_{j\in[p]} ||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_2 &= \max_{j\in[p]} ||\widehat{C}_{j,\cdot}/\widehat{\tau}_j^2 - C_{j,\cdot}/\tau_j^2||_2\\
&\leq \max_{j\in[p]} ||\widehat{\gamma}_j - \gamma_j||_2 \max_{j\in[p]} 1/\widehat{\tau}_j^2 + \max_{j\in[p]} ||\gamma_j||_2 \max_{j\in[p]} \left|1/\widehat{\tau}_j^2 - 1/\tau_j^2\right|\\
&\leq 2C_2C_\epsilon\sqrt{s\log p/n} + C_1^2/C_2C_\epsilon\sqrt{s\log p/n}\\
&= O_p(\sqrt{s\log p/n}).
\end{aligned} \tag{181}$$

For the last claim in Lemma A.20, we employ the following decomposition.

$$\begin{aligned}
\widehat{\Theta}_{j,\cdot}\Sigma\widehat{\Theta}_{j,\cdot}^\top - \Theta_{j,j} &= (\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})\Sigma(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top + 2\Theta_{j,\cdot}\Sigma(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top\\
&:= I_1 + I_2.
\end{aligned} \tag{182}$$

For I_1, by Assumption 4.2 (2), we have

$$\begin{aligned}
\max_{j\in[p]} I_1 &\leq \sigma_{\max}(\Sigma) \max_{j\in[p]} ||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_2^2\\
&= O_p(s\log p/n).
\end{aligned} \tag{183}$$

For I_2, we have

$$\begin{aligned}
\max_{j\in[p]} I_2 &= 2\max_{j\in[p]} e_j^\top(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top\\
&= 2\max_{j\in[p]} \left|1/\widehat{\tau}_j^2 - 1/\tau_j^2\right|\\
&= O_p(\sqrt{s\log p/n}).
\end{aligned} \tag{184}$$

This completes the proof of Lemma A.20.

Lemma A.21. Under Assumption 4.2, we have

$$\begin{aligned}\max_{j\in[p]}|\widehat{\sigma}_j-\sigma_j| &= O_p(s\sqrt{\log p/n}),\\ \max_{j\in[p]}|1/\widehat{\sigma}_j-1/\sigma_j| &= O_p(s\sqrt{\log p/n}).\end{aligned} \tag{185}$$

Proof of Lemma A.21. By Assumption 4.2 (2), $\sigma^2=\Theta_{j,j}\geq\sigma_{\min}(\Theta)\geq 1/C_2>0$. Therefore, we only need to consider the first claim in the lemma. In addition, since

$$\max_{j\in[p]}|\widehat{\sigma}_j-\sigma_j|\leq\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|/\sigma_j\leq\sqrt{C_2}\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|, \tag{186}$$

it is sufficient to show that

$$\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|=O_p(s\sqrt{\log p/n}). \tag{187}$$

We note that

$$\max_{j\in[p]}|\widehat{\sigma}_j^2-\sigma_j^2|\leq\max_{j\in[p]}|\widehat{\Theta}_{j,\cdot}\Sigma\widehat{\Theta}_{j,\cdot}^\top-\Theta_{j,j}|+\max_{j\in[p]}|\widehat{\Theta}_{j,\cdot}\widehat{\Sigma}\widehat{\Theta}_{j,\cdot}^\top-\widehat{\Theta}_{j,\cdot}\Sigma\widehat{\Theta}_{j,\cdot}^\top|. \tag{188}$$

By Lemma A.20, the first term is $O_p(\sqrt{s\log p/n})$. Using the same arguments as in the proof of Theorem 3.1 in Van de Geer et al. (2014) (page 1198), we can show that the second term is $O_p(s\sqrt{\log p/n})$. This completes the proof of Lemma A.21.

A.7.2 Proof of Proposition 4.2

We first note that for $j\in[p]$, the bias term Δ_j can be decomposed into the following three terms,

$$\begin{aligned}R_{1,j}&=\frac{\sqrt{n}}{\sigma_j}\Theta_{j,\cdot}P_n\dot{\ell}_{\beta^\star}-\frac{\sqrt{n}}{\widehat{\sigma}_j}\widehat{\Theta}_{j,\cdot}P_n\dot{\ell}_{\beta^\star},\\ R_{2,j}&=-\frac{\sqrt{n}}{\widehat{\sigma}_j}\left(\widehat{\Theta}_{j,\cdot}P_n\ddot{\ell}_{\widehat{\beta}}-e_j^\top\right)(\widehat{\beta}-\beta^\star),\\ R_{3,j}&=-\frac{\sqrt{n}}{\widehat{\sigma}_j}\widehat{\Theta}_{j,\cdot}P_n(\dot{\ell}_{\widehat{\beta}}-\dot{\ell}_{\beta^\star})+\frac{1}{\sqrt{n}\widehat{\sigma}_j}\widehat{\Theta}_{j,\cdot}X^\top W_{\widehat{\beta}}^2X(\widehat{\beta}-\beta^\star).\end{aligned} \tag{189}$$

We proceed to bound each term. For any $\epsilon>0$, there exists a constant $C_\epsilon>0$ such that $\mathbb{P}(\cap_{i=1}^{10}E_i^\epsilon)\geq 1-\epsilon$, in which E_i^ϵ for $i\in[4]$ are defined in Lemma A.18, E_5^ϵ and E_6^ϵ are defined in Lemma A.20, and $E_7^\epsilon, E_8^\epsilon, E_9^\epsilon, E_{10}^\epsilon$ are defined as follows,

$$\begin{aligned}E_7^\epsilon&=\left\{\max_{j\in[p]}||\widehat{\Theta}_{j,\cdot}-\Theta_{j,\cdot}||_1\leq C_\epsilon s\sqrt{\log p/n}\right\},\\ E_8^\epsilon&=\left\{\max_{j\in[p]}||\widehat{\Theta}_{j,\cdot}-\Theta_{j,\cdot}||_2\leq C_\epsilon\sqrt{s\log p/n}\right\},\\ E_9^\epsilon&=\left\{\max_{j\in[p]}|\widehat{\sigma}_j-\sigma_j|\leq C_\epsilon s\sqrt{\log p/n}\right\},\\ E_{10}^\epsilon&=\left\{\max_{j\in[p]}|1/\widehat{\sigma}_j-1/\sigma_j|\leq C_\epsilon s\sqrt{\log p/n}\right\}.\end{aligned} \tag{190}$$

In the following, we condition on this high probability event $\cap_{i=1}^{10}E_i^\epsilon$. By Assumption 4.2 (1), we further assume that n,p are large enough so that $C_\epsilon s\sqrt{\log p/n}\vee C_\epsilon\sqrt{s\log p/n}\leq 1$.

For $R_{1,j}$, we have

$$\begin{aligned}\max_{j\in[p]}R_{1,j}&\leq\max_{j\in[p]}\left|\frac{\sqrt{n}}{\sigma_j}\left(\Theta_{j,\cdot}-\widehat{\Theta}_{j,\cdot}\right)P_n\dot{\ell}_{\beta^\star}\right|+\max_{j\in[p]}\left|\sqrt{n}\widehat{\Theta}_{j,\cdot}P_n\dot{\ell}_{\beta^\star}\left(\frac{1}{\widehat{\sigma}_j}-\frac{1}{\sigma_j}\right)\right|\\ &:=I_1+I_2.\end{aligned} \tag{191}$$

For I_1, since $\mathbb{E}[P_n\dot{\ell}_{\beta^\star}] = 0$ and $||\dot{\ell}_{\beta^\star}||_\infty$ is upper bounded by Assumption 4.2 (2)(3), we have

$$||P_n\dot{\ell}_{\beta^\star}||_\infty = O_p(\sqrt{\log p/n}) \tag{192}$$

using the union bound and the Hoeffding's inequality. Since $\sigma^2 \geq 1/C_2$ (see the proof of Lemma A.21), it follows that

$$\begin{aligned} I_1 &\leq \sqrt{C_2}\sqrt{n}||P_n\dot{\ell}_{\beta^\star}||_\infty \max_{j\in[p]}||\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot}||_1 \\ &= \sqrt{n}O_p(\sqrt{\log p/n})O_p(s\sqrt{\log p/n}) \\ &= o_p(1) \end{aligned} \tag{193}$$

by Assumption 4.2 (1). For I_2, using the same argument, we have

$$\begin{aligned} I_2 &\leq \sqrt{n}\max_{j\in[p]}||\widehat{\Theta}_j||_\infty||P_n\dot{\ell}_{\beta^\star}||_\infty \max_{j\in[p]}\left|\frac{1}{\widehat{\sigma}_j} - \frac{1}{\sigma_j}\right| \\ &\leq \sqrt{n}\left(||\widehat{\Theta} - \Theta||_\infty + ||\Theta||_\infty\right)O_p(\sqrt{\log p/n})O_p(s\sqrt{\log p/n}) \\ &= o_p(1). \end{aligned} \tag{194}$$

For $R_{2,j}$, using the KKT condition of the nodewise Lasso regression, we have

$$\left|\left|\widehat{\Theta}_{j,\cdot}P_n\ddot{\ell}_{\widehat{\beta}} - e_j^\top\right|\right|_\infty \leq \lambda_j/\widehat{\tau}_j^2. \tag{195}$$

Since

$$\begin{aligned} \max_{j\in[p]} 1/\widehat{\tau}_j^2 &\leq \max_{j\in[p]}|1/\widehat{\tau}_j^2 - 1/\tau_j^2| + \max_{j\in[p]} 1/\tau_j^2 \\ &\leq C_\epsilon\sqrt{s\log p/n} + \max_{j\in[p]} 1/\Theta_{j,j} \\ &\leq 1 + C_2 \end{aligned} \tag{196}$$

and

$$\begin{aligned} \max_{j\in[p]} 1/\widehat{\sigma}_j &\leq \max_{j\in[p]}|1/\widehat{\sigma}_j - 1/\sigma_j| + \max_{j\in[p]} 1/\sigma_j \\ &\leq C_\epsilon s\sqrt{\log p/n} + 1/\sigma_{\min}^{1/2}(\Sigma) \\ &\leq 1 + \sqrt{C_2} \end{aligned} \tag{197}$$

by Assumption 4.2 (3), we have

$$R_{2,j} \leq (1 + C_2)\sqrt{n}\left|\left|\widehat{\Theta}_{j,\cdot}P_n\ddot{\ell}_{\widehat{\beta}} - e_j^\top\right|\right|_\infty ||\widehat{\beta} - \beta^\star||_1 = o_p(1) \tag{198}$$

by Assumption 4.2 (1).

For $R_{3,j}$, the first term is $o_p(1)$ following the proof of Theorem 3.1 in Van de Geer et al. (2014). For the second term, we note that

$$\begin{aligned} ||X\widehat{\Theta}_{j,\cdot}^\top||_\infty &\leq ||X\Theta_{j,\cdot}^\top||_\infty + ||X(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top||_\infty \\ &\leq ||X||_\infty + ||X_{-j}\gamma_j||_\infty + ||X||_\infty||(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top||_1 \\ &\leq ||X||_\infty + ||X_{\beta^\star,-j}\gamma_j||_\infty/\inf|\ddot{\rho}(x^\intercal\beta^\star)| + ||X||_\infty||(\widehat{\Theta}_{j,\cdot} - \Theta_{j,\cdot})^\top||_1 \\ &< +\infty \end{aligned} \tag{199}$$

by Assumption 4.2 (2)(3). By the mean value Theorem and Assumption 4.2, it follows that

$$R_{3,j} \leq \frac{1}{\sqrt{n}\widehat{\sigma}_j}||X\widehat{\Theta}_{j,\cdot}^\top||_\infty||X(\widehat{\beta} - \beta^\star)||_2^2 = o_p(1). \tag{200}$$

This completes the proof of Proposition 4.2.

A.7.3 Proof of Proposition 4.3

The proof of Proposition 4.3 essentially follows the same as the proof of Proposition 4.1. The only change is that $Z_j = -\sqrt{n}\Theta_{j,\cdot}P_n\dot{\ell}_{\beta^\star}/\sigma_j$ is not exactly normal, but asymptotically normal. The discrepancy between the joint law of $(Z_1, \cdots, Z_p)$ and the corresponding multivariate normal distribution (with the covariance matrix Σ) can be quantified using the Berry-Esseen Theorem, which is in the order of $o(1/\sqrt{n})$. Thus, we can establish the GLM version of Lemma A.15 similarly.

B Additional Numerical Results

B.1 The moderate-dimensional setting

B.1.1 Logistic regression

We provide additional numerical results to complement Section 5.1.1 by considering the following types of the covariance matrix Σ.

1. The small-n-and-p setting.
   (a) Features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i\neq j)}$.
   (b) Features have constant pairwise partial correlation, i.e., $\Sigma^{-1}_{ij} = r^{\mathbb{1}(i\neq j)}$.
   (c) Features have Toeplitz partial correlation, i.e., $\Sigma^{-1}_{ij} = r^{|i-j|}$.
2. The large-n-and-p setting.
   (a) Features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i\neq j)}$.

For the small-n-and-p setting, the empirical FDRs and powers of different methods in scenarios (a), (b) and (c) are summarized in Figures 8, 9 and 10, respectively. For the large-n-and-p setting, the results in scenario (a) are shown in Figure 11.

B.1.2 Negative binomial regression

To complement Section 5.1.2, we provide additional numerical results for the case where features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i\neq j)}$. The empirical FDRs and powers of different methods are summarized in Figure 12.

B.2 The high-dimensional setting

B.2.1 Linear regression

To complement Section 5.2.1, we first detail the Toeplitz matrix aligned along the diagonal of the covariance matrix Σ as below,

$$\begin{bmatrix} 1 & \frac{(p'-2)r}{p'-1} & \frac{(p'-3)r}{p'-1} & \cdots & \frac{r}{p'-1} & 0 \\ \frac{(p'-2)r}{p'-1} & 1 & \frac{(p'-2)r}{p'-1} & \cdots & \frac{2r}{p'-1} & \frac{r}{p'-1} \\ \vdots & & \ddots & & & \vdots \\ 0 & \frac{r}{p'-1} & \frac{2r}{p'-1} & \cdots & \frac{(p'-2)r}{p'-1} & 1 \end{bmatrix}, \tag{201}$$

where $p' = p/10$. Throughout, we refer $r \in (0, 1)$ as the correlation factor.

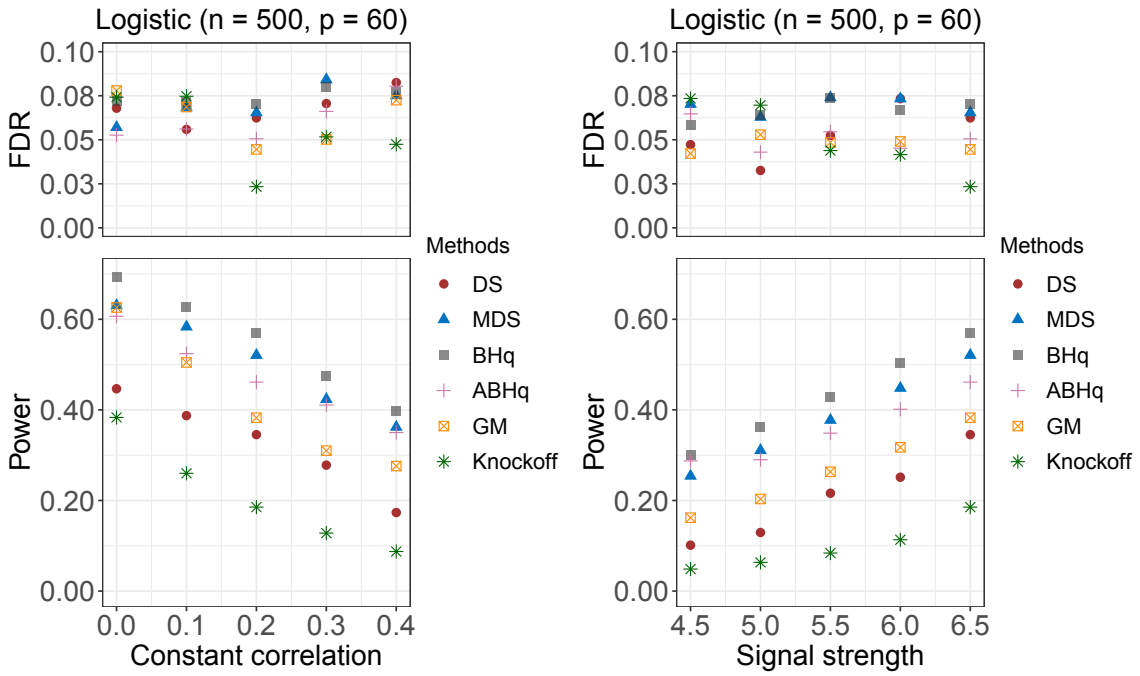

Figure 8: Empirical FDRs and powers for the logistic regression model in the small-n-and-p setting. The algorithmic settings are as per Figure 2, except that features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i \neq j)}$.

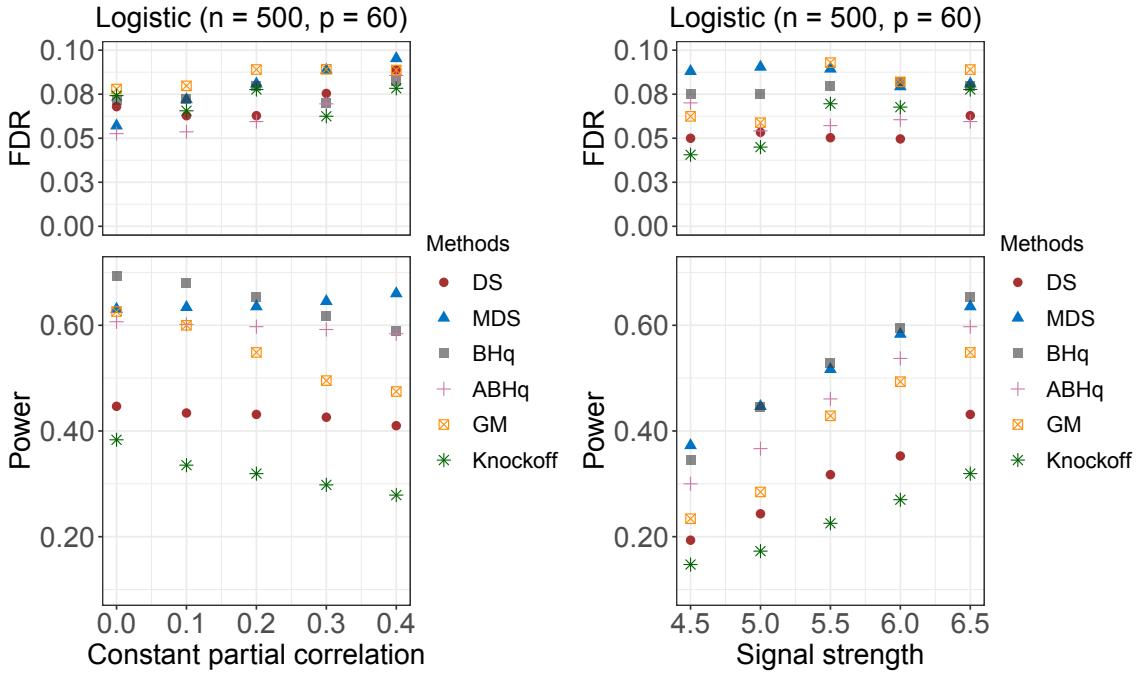

Figure 9: Empirical FDRs and powers for the logistic regression model in the small-n-and-p setting. The algorithmic settings are as per Figure 2, except that features have constant pairwise partial correlation, i.e., $\Sigma^{-1}_{ij} = r^{\mathbb{1}(i \neq j)}$.

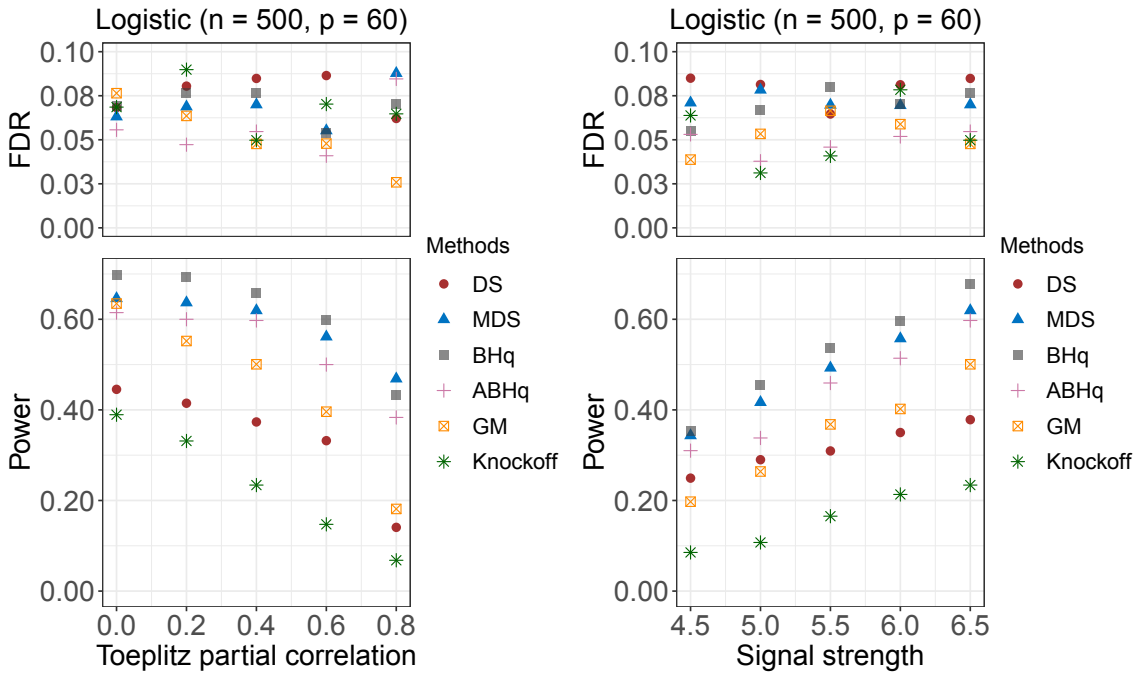

Figure 10: Empirical FDRs and powers for the logistic regression model in the small-n-and-p setting. The algorithmic settings are as per Figure 2, except that features have Toeplitz partial correlation, i.e., $\Sigma^{-1}_{ij} = r^{|i-j|}$.

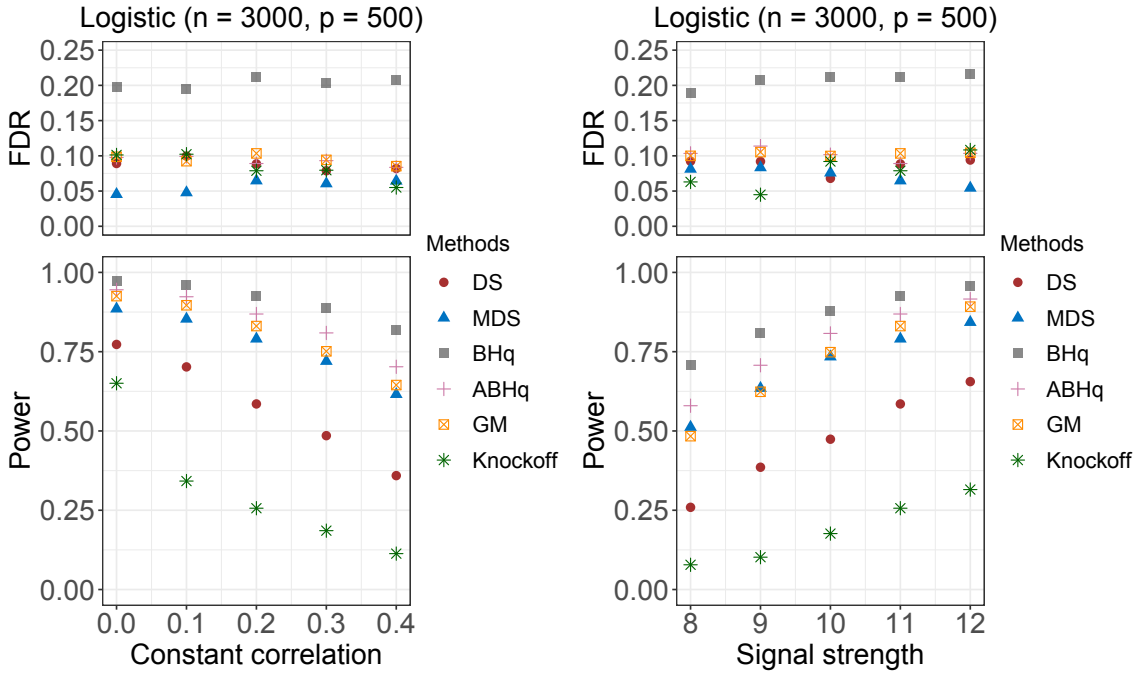

Figure 11: Empirical FDRs and powers for the logistic regression model in the large-n-and-p setting. The algorithmic settings are as per Figure 3, except that features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i \neq j)}$.

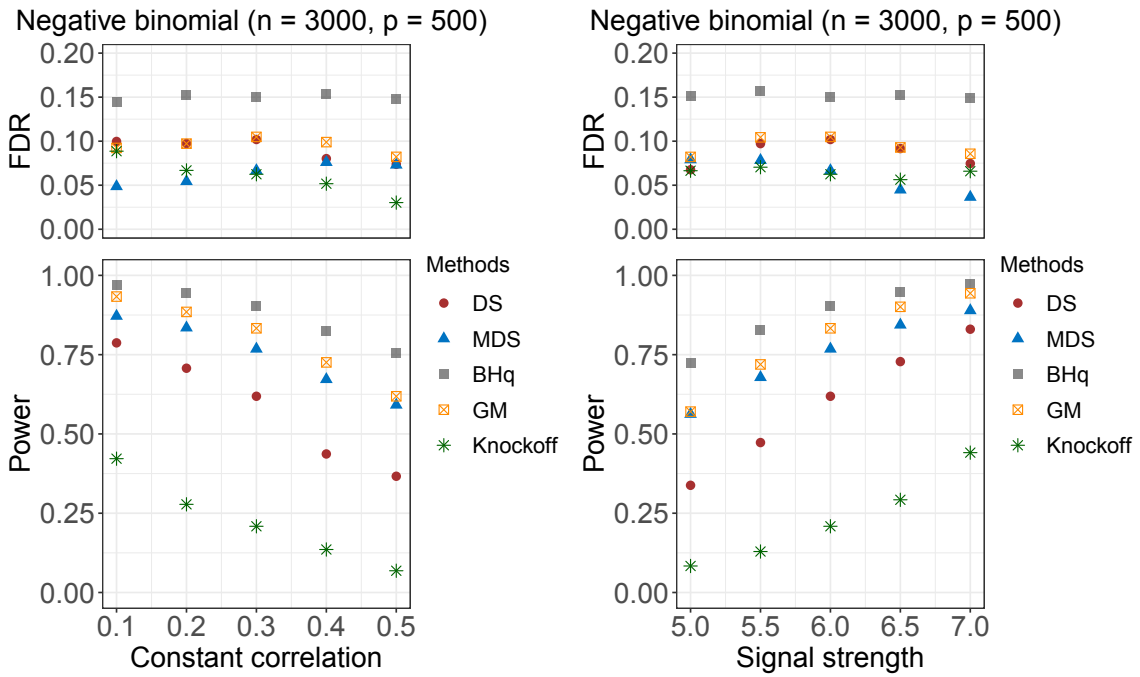

Figure 12: Empirical FDRs and powers for the negative binomial regression model. The algorithmic settings are as per Figure 4, except that features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i \neq j)}$.

We then provide additional numerical results for the case where features have constant pairwise correlation, i.e., $\Sigma_{ij} = r^{\mathbb{1}(i \neq j)}$. All the algorithmic settings are the same as discussed in Section 5.2.1, except that the number of relevant features is 50 across all settings, i.e. $p_1 = 50$. The empirical FDRs and powers of different methods are summarized in Figure 13.

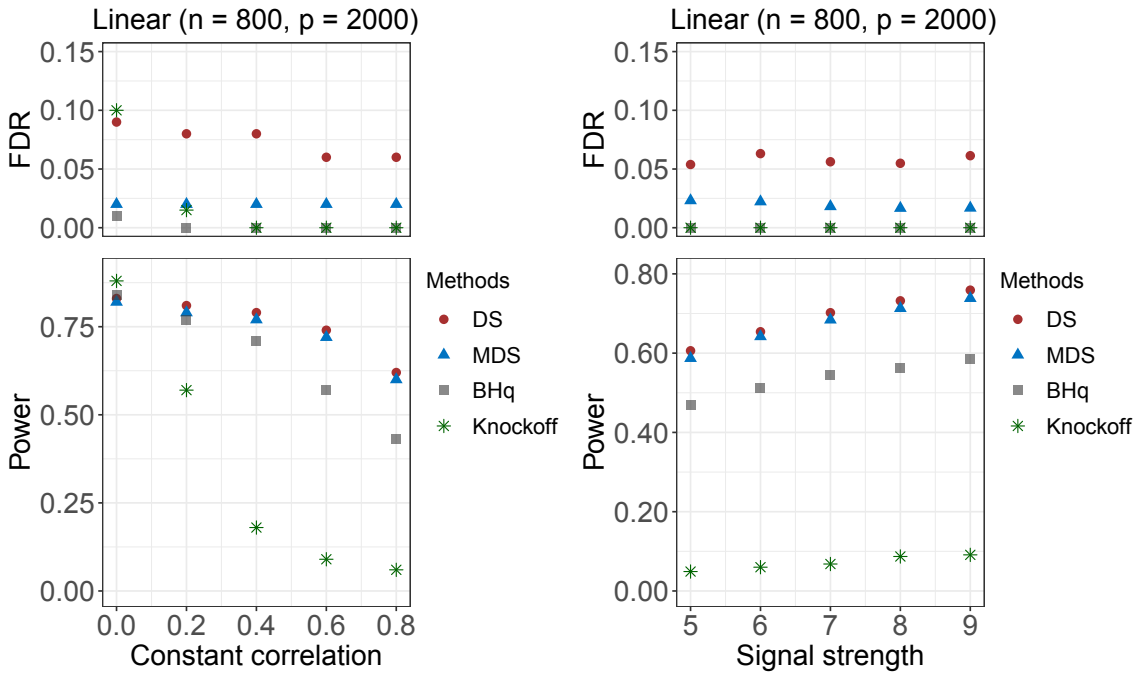

Figure 13: Empirical FDRs and powers for the linear model. Each row of the design matrix is independently drawn from $N(0, \Sigma)$, with $\Sigma_{ij} = r^{\mathbb{1}(i \neq j)}$, i.e., features have constant pairwise correlation. $\beta_j^\star$ for $j \in S_1$ are i.i.d. samples from $N(0, s^2)$, where s is referred to as the signal strength. The signal strength along the x-axis of the right panel shows multiples of $\sqrt{\log p/n}$. In the left panel, we fix the signal strength at $8\sqrt{\log p/n}$ and vary the correlation r. In the right panel, we fix the ccorrelation at $r = 0.6$ and vary the signal strength. The number of relevant features is 50 across all settings, and the designated FDR control level is $q = 0.1$. Each dot in the figure represents the average from 50 independent runs.