Taylor & Francis
Taylor & Francis Group

Check for updates

RANK: Large-Scale Inference With Graphical Nonlinear Knockoffs

Yingying Fan[a], Emre Demirkaya[b], Gaorong Li[c], and Jinchi Lv[a]

[a]Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA; [b]Department of Business Analytics and Statistics, Haslam College of Business, University of Tennessee, Knoxville, TN; [c]Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing, China

ABSTRACT

Power and reproducibility are key to enabling refined scientific discoveries in contemporary big data applications with general high-dimensional nonlinear models. In this article, we provide theoretical foundations on the power and robustness for the model-X knockoffs procedure introduced recently in Candès, Fan, Janson and Lv in high-dimensional setting when the covariate distribution is characterized by Gaussian graphical model. We establish that under mild regularity conditions, the power of the oracle knockoffs procedure with known covariate distribution in high-dimensional linear models is asymptotically one as sample size goes to infinity. When moving away from the ideal case, we suggest the modified model-X knockoffs method called graphical nonlinear knockoffs (RANK) to accommodate the unknown covariate distribution. We provide theoretical justifications on the robustness of our modified procedure by showing that the false discovery rate (FDR) is asymptotically controlled at the target level and the power is asymptotically one with the estimated covariate distribution. To the best of our knowledge, this is the first formal theoretical result on the power for the knockoffs procedure. Simulation results demonstrate that compared to existing approaches, our method performs competitively in both FDR control and power. A real dataset is analyzed to further assess the performance of the suggested knockoffs procedure. Supplementary materials for this article are available online.

1. Introduction

Feature selection with big data is of fundamental importance to many contemporary applications from different disciplines of social sciences, health sciences, and engineering Hastie, Tibshirani, and Friedman (2009), Fan and Lv (2010), and Bühlmann and van de Geer (2011). Over the past two decades, various feature selection methods, theory, and algorithms have been extensively developed and investigated for a wide spectrum of flexible models ranging from parametric to semiparametric and nonparametric linking a high-dimensional covariate vector $\mathbf{x} = (X_1, \ldots, X_p)^T$ of p features X_j's to a response Y of interest, where the dimensionality p can be large compared to the available sample size n or even greatly exceed n. The success of feature selection for enhanced prediction in practice can be attributed to the reduction of noise accumulation associated with high-dimensional data through dimensionality reduction. In particular, most existing studies have focused on the power perspective of feature selection procedures such as the sure screening property, model selection consistency, oracle property, and oracle inequalities. When the model is correctly specified, researchers and practitioners often would like to know whether the estimated model involving a subset of the p covariates enjoys reproducibility in that the fraction of noise features in the discovered model is controlled. Yet such a practical issue of reproducibility is largely less well understood for the settings of general high-dimensional nonlinear models.

Moreover, it is no longer clear whether the power of feature selection procedures can be retained when one intends to ensure the reproducibility.

Indeed, the issues of power and reproducibility are key to enabling refined scientific discoveries in big data applications utilizing general high-dimensional nonlinear models. To characterize the reproducibility of statistical inference, the seminal paper of Benjamini and Hochberg (1995) introduced an elegant concept of false discovery rate (FDR), which is defined as the expectation of the fraction of false discoveries among all the discoveries, and proposed a popularly used Benjamini–Hochberg procedure for FDR control by resorting to the p-values for large-scale multiple testing returned by some statistical estimation and testing procedure. There is a huge literature on FDR control for large-scale inference and various generalizations and extensions of the original FDR procedure were developed and investigated for different settings and applications Benjamini and Yekutieli (2001), Efron and Tibshirani (2002), Storey (2002), Storey, Taylor, and Siegmund (2004), Abramovich et al. (2006), Efron (2007a, 2007b), Fan, Hall, and Yao (2007), Wu (2008), Clarke and Hall (2009), Hall and Wang (2010), Fan and Fan (2011), Meng et al. (2011), Zhang and Liu (2011), Fan, Han, and Gu (2012), Liu and Shao (2014), and Su and Candès (2016). Most of existing work either assumes a specific functional form such as linearity on the dependence structure of response Y on covariates X_j's, or relies on the p-values for evaluating the

significance of covariates X_j's. Yet in high-dimensional settings, we often do not have such luxury since response Y could depend on covariate vector $\mathbf{x}$ through very complicated forms and even when Y and $\mathbf{x}$ have simple dependence structure, high dimensionality of covariates can render classical p-value calculation procedures no longer justified or simply invalid Huber (1973), Fan, Demirkaya, and Lv (2017), and Sur, Chen, and Candès (2017). These intrinsic challenges can make the p-value based methods difficult to apply or even fail Candès et al. (2018).

To accommodate arbitrary dependence structure of Y on $\mathbf{x}$ and bypass the need of calculating accurate p-values for covariate significance, Candès et al. (2018) recently introduced the model-X knockoffs framework for FDR control in general high-dimensional nonlinear models. Their work was inspired by and builds upon the ingenious development of the knockoff filter in Barber and Candès (2015), which provides effective FDR control in the setting of Gaussian linear model with dimensionality p no larger than sample size n. The knockoff filter was later extended in Barber and Candès (2016) to high-dimensional linear model using the ideas of data splitting and feature screening. The salient idea of Barber and Candès (2015) is to construct the so-called "knockoff" variables which mimic the dependence structure of the original covariates but are independent of response Y conditional on the original covariates. These knockoff variables can be used as control variables. By comparing the regression outcomes for original variables with those for control variables, the relevant set of variables can be identified more accurately and thus the FDR can be better controlled. The model-X knockoffs framework introduced in Candès et al. (2018) greatly expands the applicability of the original knockoff filter in that the response Y and covariates $\mathbf{x}$ can have arbitrarily complicated dependence structure and the dimensionality p can be arbitrarily large compared to sample size n. It was theoretically justified in Candès et al. (2018) that the model-X knockoffs procedure controls FDR exactly in finite samples of arbitrary dimensions. However, one important assumption in their theoretical development is that the joint distribution of covariates $\mathbf{x}$ should be known. Moreover, formal power analysis of the knockoffs framework is still lacking even for the setting of Gaussian linear model.

Despite the importance of known covariate distribution in their theoretical development, Candès et al. (2018) empirically explored the scenario of unknown covariate distribution for the specific setting of generalized linear model (GLM) McCullagh and Nelder (1989) with Gaussian design matrix and discovered that the estimation error of the covariate distribution can have negligible effect on FDR control. Yet there exist no formal theoretical justifications on the robustness of the model-X knockoffs method and it is also unclear to what extent such robustness can hold beyond the GLM setting. To address these fundamental challenges, our article intends as the first attempt to provide theoretical foundations on the power and robustness for the model-X knockoffs framework. Specifically, the major innovations of the article are 2-fold. First, we will formally investigate the power of the knockoffs framework in high-dimensional linear models with both known and unknown covariate distribution. Second, we will provide theoretical support on the robustness of the model-X knockoffs procedure with unknown covariate distribution in general high-dimensional nonlinear models.

More specifically, in the ideal case of known covariate distribution, we prove that the model-X knockoffs procedure in Candès et al. (2018) has asymptotic power one under mild regularity conditions in high-dimensional linear models. When moving away from the ideal scenario, to accommodate the difficulty caused by unknown covariate distribution we suggest the modified model-X knockoffs method called graphical nonlinear knockoffs (RANK). The modified knockoffs procedure exploits the data splitting idea, where the first half of the sample is used to estimate the unknown covariate distribution and reduce the model size, and the second half of the sample is employed to *globally* construct the knockoff variables and apply the knockoffs procedure. We establish that the modified knockoffs procedure asymptotically controls the FDR regardless of whether the reduced model contains the true model or not. Such feature makes our work intrinsically different from that in Barber and Candès (2016) requiring the sure screening property Fan and Lv (2008) of the reduced model; see Section 3.1 for more detailed discussions on the differences. In our theoretical analysis of FDR, we still allow for arbitrary dependence structure of response Y on covariates $\mathbf{x}$ and assume that the joint distribution of $\mathbf{x}$ is characterized by Gaussian graphical model with unknown precision matrix Lauritzen (1996). In the specific case of high-dimensional linear models with unknown covariate distribution, we also provide robustness analysis on the power of our modified procedure.

The rest of the article is organized as follows. Section 2 reviews the model-X knockoffs framework and provides theoretical justifications on its power in high-dimensional linear models. We introduce the modified model-X knockoffs procedure RANK and investigate its robustness on both FDR control and power with respect to the estimation of unknown covariate distribution in Section 3. Section 4 presents several simulation examples of both linear and nonlinear models to verify our theoretical results. We demonstrate the performance of our procedure on a real dataset in Section 5. Section 6 discusses some implications and extensions of our work. The proofs of main results are relegated to the Appendix. Additional technical details are provided in the supplementary materials.

2. Power Analysis for Oracle Model-X Knockoffs

Suppose we have a sample $(\mathbf{x}_i, Y_i)_{i=1}^n$ of n independent and identically distributed (iid) observations from the population $(\mathbf{x}, Y)$, where dimensionality p of covariate vector $\mathbf{x} = (X_1, \ldots, X_p)^T$ can greatly exceed available sample size n. To ensure model identifiability, it is common to assume that only a small fraction of p covariates X_j's are truly relevant to response Y. To be more precise, Candès et al. (2018) defined the set of irrelevant features $\mathcal{S}_1$ as that consisting of X_j's such that X_j is independent of Y conditional on all remaining $p - 1$ covariates X_k's with $k \neq j$, and thus the set of truly relevant features $\mathcal{S}_0$ is given naturally by $\mathcal{S}_1^c$, the complement of set $\mathcal{S}_1$. Features in sets $\mathcal{S}_0$ and $\mathcal{S}_0^c = \mathcal{S}_1$ are also referred to as important and noise features, respectively.

We aim at accurately identifying these truly relevant features in set $\mathcal{S}_0$, that is, assumed to be identifiable while keeping the FDR Benjamini and Hochberg (1995) under control. The FDR

for a feature selection procedure is defined as

$$\text{FDR} = \mathbb{E}[\text{FDP}] \quad \text{with FDP} = \frac{|\widehat{\mathcal{S}} \cap \mathcal{S}_0^c|}{|\widehat{\mathcal{S}}|}, \tag{1}$$

where $\widehat{\mathcal{S}}$ denotes the sparse model returned by the feature selection procedure, $|\cdot|$ stands for the cardinality of a set, and the convention $0/0 = 0$ is used in the definition of the false discovery proportion (FDP), which is the fraction of noise features in the discovered model. Here feature selection procedure can be any favorite sparse modeling method by the choice of the user.

2.1. Review of Model-X Knockoffs Framework

Our suggested graphical nonlinear knockoffs procedure in Section 3 falls in the general framework of model-X knockoffs introduced in Candès et al. (2018), which we briefly review in this section. The key ingredient of model-X knockoffs framework is the construction of the so-called model-X knockoff variables that are defined as follows.

Definition 1 (Candès et al. 2018). Model-X knockoffs for the family of random variables $\mathbf{x} = (X_1, \ldots, X_p)^T$ is a new family of random variables $\widetilde{\mathbf{x}} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)^T$ that satisfies two properties: (1) $(\mathbf{x}^T, \widetilde{\mathbf{x}}^T)_{\text{swap}(\mathcal{S})} \overset{d}{=} (\mathbf{x}^T, \widetilde{\mathbf{x}}^T)$ for any subset $\mathcal{S} \subset \{1, \ldots, p\}$, where swap($\mathcal{S}$) means swapping components X_j and $\widetilde{X}_j$ for each $j \in \mathcal{S}$ and $\overset{d}{=}$ denotes equal in distribution, and (2) $\widetilde{\mathbf{x}} \perp\!\!\!\perp Y | \mathbf{x}$.

We see from Definition 1 that model-X knockoff variables $\widetilde{X}_j$'s mimic the probabilistic dependency structure among the original features X_j's and are independent of response Y given X_j's. When the covariate distribution is characterized by Gaussian graphical model Lauritzen (1996), that is,

$$\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Omega}_0^{-1}) \tag{2}$$

with $p \times p$ precision matrix $\mathbf{\Omega}_0$ encoding the graphical structure of the conditional dependency among the covariates X_j's, we can construct the p-variate model-X knockoff random variable $\widetilde{\mathbf{x}}$ characterized in Definition 1 as

$$\widetilde{\mathbf{x}}|\mathbf{x} \sim N\Big(\mathbf{x} - \text{diag}\{\mathbf{s}\}\mathbf{\Omega}_0\mathbf{x}, 2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\mathbf{\Omega}_0\text{diag}\{\mathbf{s}\}\Big), \tag{3}$$

where $\mathbf{s}$ is a p-dimensional vector with nonnegative components chosen in a suitable way. In fact, in view of Equations (2) and (3) it is easy to show that the original features and model-X knockoff variables have the following joint distribution

$$\begin{pmatrix} \mathbf{x} \\ \widetilde{\mathbf{x}} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_0 & \mathbf{\Sigma}_0 - \text{diag}\{\mathbf{s}\} \\ \mathbf{\Sigma}_0 - \text{diag}\{\mathbf{s}\} & \mathbf{\Sigma}_0 \end{pmatrix}\right) \tag{4}$$

with $\mathbf{\Sigma}_0 = \mathbf{\Omega}_0^{-1}$ the covariance matrix of covariates $\mathbf{x}$. Intuitively, larger components of $\mathbf{s}$ means that the constructed knockoff variables deviate further from the original features, resulting in higher power in distinguishing them. The p-dimensional vector $\mathbf{s}$ in Equation (3) should be chosen in a way such that $\mathbf{\Sigma}_0 - 2^{-1}\text{diag}\{\mathbf{s}\}$ is positive definite, and can be selected using the methods in Candès et al. (2018). We will treat it as a nuisance parameter throughout our theoretical analysis.

With the constructed knockoff variables $\widetilde{\mathbf{x}}$, the knockoffs inference framework proceeds as follows. We select important variables by resorting to the knockoff statistics $W_j = f_j(Z_j, \widetilde{Z}_j)$ defined for each $1 \leq j \leq p$, where Z_j and $\widetilde{Z}_j$ represent feature importance measures for jth covariate X_j and its knockoff counterpart $\widetilde{X}_j$, respectively, and $f_j(\cdot, \cdot)$ is an antisymmetric function satisfying $f_j(z_j, \tilde{z}_j) = -f_j(\tilde{z}_j, z_j)$. For example, in linear regression models, one can choose Z_j and $\widetilde{Z}_j$ as the Lasso Tibshirani (1996) regression coefficients of X_j and $\widetilde{X}_j$, respectively, and a valid knockoff statistic is $W_j = f_j(z_j, \tilde{z}_j) = |z_j| - |\tilde{z}_j|$. There are also many other options for defining the feature importance measures. Observe that all model-X knockoff variables $\widetilde{X}_j$'s are just noise features by the second property in Definition 1. Thus, intuitively, a large positive value of knockoff statistic W_j indicates that jth covariate X_j is important, while a small magnitude of W_j usually corresponds to noise features.

The final step of the knockoffs inference framework is to sort $|W_j|$'s from high to low and select features whose W_j's are at or above some threshold T, which results in the discovered model

$$\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(T) = \{1 \leq j \leq p : W_j \geq T\}. \tag{5}$$

Following Barber and Candès (2015) and Candès et al. (2018), one can choose the threshold T in the following two ways

$$T = \min\left\{t \in \mathcal{W} : \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q\right\}, \tag{6}$$

$$T_+ = \min\left\{t \in \mathcal{W} : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}|} \leq q\right\}, \tag{7}$$

where $\mathcal{W} = \{|W_j| : 1 \leq j \leq p\} \setminus \{0\}$ is the set of unique nonzero values attained by $|W_j|$'s and $q \in (0, 1)$ is the desired FDR level specified by the user. The procedures using threshold T in Equation (6) and threshold T_+ in Equation (7) are referred to as knockoffs and knockoffs$_+$ methods, respectively. It was proved in Candès et al. (2018) that model-X knockoffs procedure controls a modified FDR that replaces $|\widehat{\mathcal{S}}|$ in the denominator by $q^{-1} + |\widehat{\mathcal{S}}|$ in Equation (1), and model-X knockoffs$_+$ procedure achieves exact FDR control in finite samples regardless of dimensionality p and dependence structure of response Y on covariates $\mathbf{x}$. The major assumption needed in Candès et al. (2018) is that the distribution of covariates $\mathbf{x}$ is known. Throughout the article, we implicitly use the threshold T_+ defined in Equation (7) for FDR control in the knockoffs inference framework but still write it as T for notational simplicity.

2.2. Power Analysis in Linear Models

Although the knockoffs procedures were proved rigorously to have controlled FDR in Barber and Candès (2015), Barber and Candès (2016), and Candès et al. (2018), their power advantages over popularly used approaches have been demonstrated only numerically therein. In fact, formal power analysis for the knockoffs framework is still lacking even in simple model settings such as linear regression. We aim to fill in this gap as a first attempt and provide theoretical foundations on the power analysis for model-X knockoffs framework. In this section, we will focus on the *oracle* model-X knockoffs procedure for the

ideal case when the true precision matrix $\boldsymbol{\Omega}_0$ for the covariate distribution in Equation (2) is known, which is the setting assumed in Candès et al. (2018). The robustness analysis for the case of unknown precision matrix $\boldsymbol{\Omega}_0$ will be undertaken in Section 3.

We would like to remark that the power analysis for the knockoffs framework is necessary and nontrivial. The FDR and power are two sides of the same coin, just like Type I and Type II errors in hypothesis testing. The knockoffs framework is a wrapper and can be combined with most model selection methods to achieve FDR control. Yet the theoretical properties of power after applying the knockoffs procedure are completely unknown for the case of correlated covariates and unknown covariate distribution. For example, when the knockoffs framework is combined with the Lasso, it further selects variables from the set of variables picked by Lasso applied with the augmented design matrix to achieve the FDR control. For this reason, the power of knockoffs is usually lower than that of Lasso. The main focus of this section is to investigate how much power loss the knockoffs framework would encounter when combined with Lasso.

Since the power analysis for the knockoffs framework is nontrivial and challenging, we content ourselves on the setting of high-dimensional linear models for the technical analysis on power. The linear regression model assumes that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \qquad (8)$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ is an n-dimensional response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is an $n \times p$ design matrix consisting of p covariates X_j's, $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ is a p-dimensional true regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-dimensional error vector independent of $\mathbf{X}$. As mentioned before, the true model $\mathcal{S}_0 = \text{supp}(\boldsymbol{\beta}_0)$ which is the support of $\boldsymbol{\beta}_0$ is assumed to be sparse with size $s = |\mathcal{S}_0|$, and the n rows of design matrix $\mathbf{X}$ are iid observations generated from Gaussian graphical model (2). Without loss of generality, all the diagonal entries of covariance matrix $\boldsymbol{\Sigma}_0$ are assumed to be ones.

As discussed in Section 2.1, there are many choices of the feature selection procedure up to the user for producing the feature importance measures Z_j and $\widetilde{Z}_j$ for covariates X_j and knockoff variables $\widetilde{X}_j$, respectively, and there are also different ways to construct the knockoff statistics W_j. For the illustration purpose, we adopt the Lasso coefficient difference (LCD) as the knockoff statistics in our power analysis. The specific choice of LCD for knockoff statistics was proposed and recommended in Candès et al. (2018), in which it was demonstrated empirically to outperform some other choices in terms of power. The LCD is formally defined as

$$W_j = |\widehat{\beta}_j(\lambda)| - |\widehat{\beta}_{p+j}(\lambda)|, \qquad (9)$$

where $\widehat{\beta}_j(\lambda)$ and $\widehat{\beta}_{p+j}(\lambda)$ denote the jth and $(p+j)$th components, respectively, of the Lasso Tibshirani (1996) regression coefficient vector

$$\widehat{\boldsymbol{\beta}}(\lambda) = \text{argmin}_{\mathbf{b} \in \mathbb{R}^{2p}} \left\{ (2n)^{-1} \|\mathbf{y} - [\mathbf{X}, \widetilde{\mathbf{X}}]\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}, \qquad (10)$$

with $\lambda \geq 0$ the regularization parameter, $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n)^T$ an $n \times p$ matrix whose n rows are independent random vectors of model-X knockoff variables generated from Equation (3), and $\| \cdot \|_r$ for $r \geq 0$ the L_r-norm of a vector. To simplify the technical analysis, we assume that with asymptotic probability one, there are no ties in the magnitude of nonzero W_j's and no ties in the magnitude of nonzero components of Lasso solution in Equation (10), which is a mild condition in light of the continuity of the underlying distributions.

To facilitate the power analysis, we impose some basic regularity conditions.

Condition 1. The components of $\boldsymbol{\varepsilon}$ are iid with sub-Gaussian distribution.

Condition 2. It holds that $\min_{j \in \mathcal{S}_0} |\beta_{0,j}| \geq \kappa_n \{(\log p)/n\}^{1/2}$ for some slowly diverging sequence $\kappa_n \to \infty$ as $n \to \infty$.

Condition 3. There exists some constant $c \in (2(qs)^{-1}, 1)$ such that with asymptotic probability one, $|\widehat{\mathcal{S}}| \geq cs$ for $\widehat{\mathcal{S}}$ given in Equation (5).

Condition 1 can be relaxed to heavier-tailed distributions at the cost of slower convergence rates as long as similar concentration inequalities used in the proofs continue to hold. Condition 2 is assumed to ensure that the Lasso solution $\widehat{\boldsymbol{\beta}}(\lambda)$ does not miss a great portion of important features in $\mathcal{S}_0$. This is necessary since the knockoffs procedure under investigation builds upon the Lasso solution and thus its power is naturally upper bounded by that of Lasso. To see this, recall the well-known oracle inequality for Lasso Bickel, Ritov, and Tsybakov (2009) and Bühlmann and van de Geer (2011) that with asymptotic probability one, $\|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0\|_2 = O(s^{1/2}\lambda)$ for λ chosen in the order of $\{(\log p)/n\}^{1/2}$. Then Condition 2 entails that for some $\kappa_n \to \infty$, $O(s\lambda^2) = \|\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0\|_2^2 \geq \sum_{j \in \widehat{\mathcal{S}}_{\text{L}}^c \cap \mathcal{S}_0} \beta_{0,j}^2 \geq n^{-1}(\log p)\kappa_n^2|\widehat{\mathcal{S}}_{\text{L}}^c \cap \mathcal{S}_0|$ with $\widehat{\mathcal{S}}_{\text{L}} = \text{supp}\{\widehat{\boldsymbol{\beta}}(\lambda)\}$. Thus, the number of important features missed by Lasso $|\widehat{\mathcal{S}}_{\text{L}}^c \cap \mathcal{S}_0|$ is upper bounded by $O(s\kappa_n^{-2})$ with asymptotic probability one. This guarantees that the power of Lasso is lowered bounded by $1 - O(\kappa_n^{-2})$; that is, Lasso has asymptotic power one. However, as discussed previously the power of knockoffs is always upper bounded by that of Lasso. So we are interested in the relative power of knockoffs compared to that of Lasso. For this reason, Condition 2 is imposed to simplify the technical analysis of the knockoffs power by ensuring that the asymptotic power of Lasso is one. We will show in Theorem 1 that there is almost no power loss when applying model-X knockoffs procedure.

Condition 3 imposes a lower bound on the size of the sparse model selected by the knockoffs procedure. Recall that we assume the number of true variables s can diverge with sample size n. The rationale behind Condition 3 is that any method with high power should at least be able to select a large number of variables which are not necessarily true ones though. Since it is not straightforward to check, we provide a sufficient condition, that is, more intuitive in Lemma 1 below, which shows that Condition 3 can hold as long as there exist enough strong signals in the model. We acknowledge that Lemma 1 may not be a necessary condition for Condition 3.

Lemma 1. Assume that Condition 1 holds and there exists some constant $c \in (2(qs)^{-1}, 1)$ such that $|\mathcal{S}_2| \geq cs$ with $\mathcal{S}_2 = \{j : |\beta_{0,j}| \gg [sn^{-1}(\log p)]^{1/2}\}$. Then Condition 3 holds.

We would like to mention that the conditions of Lemma 1 are not stronger than Condition 2. We require a few strong signals, and yet still allow for many very weak ones. In other words, the set of strong signals $\mathcal{S}_2$ is only a large enough proper subset of the set of all signals $\mathcal{S}_0$.

We are now ready to characterize the statistical power of the knockoffs procedure in high-dimensional linear model (8). Formally speaking, the power of a feature selection procedure is defined as

$$\text{Power}(\widehat{\mathcal{S}}) = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}_0|}{|\mathcal{S}_0|}\right], \quad (11)$$

where $\widehat{\mathcal{S}}$ denotes the discovered sparse model returned by the feature selection procedure.

Theorem 1. Assume that Condition 1–3 hold, all the eigenvalues of $\boldsymbol{\Omega}_0$ are bounded away from 0 to ∞, the smallest eigenvalue of $2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})$ is positive and bounded away from 0, and $\lambda = C_\lambda\{(\log p)/n\}^{1/2}$ with $C_\lambda > 0$ some constant. Then the oracle model-X knockoffs procedure satisfies that with probability at least $1 - c_3 p^{-c_3}$,

$$|\widehat{\mathcal{S}} \cap \mathcal{S}_0|/|\mathcal{S}_0| \geq 1 - C_{\ell_1} C_\lambda(\varphi + 1)\kappa_n^{-1}$$

and therefore,

$$\text{Power}(\widehat{\mathcal{S}}) = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}_0|}{|\mathcal{S}_0|}\right] \geq 1 - C_{\ell_1} C_\lambda(\varphi + 1)\kappa_n^{-1} - c_3 p^{-c_3}$$
$$+ o(\kappa_n^{-1}) \to 1$$

as $n \to \infty$, where φ is the golden ratio and C_{ℓ_1} is some positive constant.

Theorem 1 reveals that the oracle model-X knockoffs procedure in Candès et al. (2018) knowing the true precision matrix $\boldsymbol{\Omega}_0$ for the covariate distribution can indeed have asymptotic power one under some mild regularity conditions. Since parameter κ_n characterizes the signal strength, it is seen that the stronger the signal, the faster the convergence of power to one. This shows that for the ideal case, model-X knockoffs procedure can enjoy appealing FDR control and power properties simultaneously.

3. Robustness of Graphical Nonlinear Knockoffs

When moving away from the ideal scenario considered in Section 2, a natural question is whether both properties of FDR control and power can continue to hold with no access to the knowledge of true covariate distribution. To gain insights into such a question, we now turn to investigating the robustness of model-X knockoffs framework. Hereafter, we assume that the true precision matrix $\boldsymbol{\Omega}_0$ for the covariate distribution in Equation (2) is *unknown*. We will begin with the FDR analysis and then move on to the power analysis.

3.1. Modified Model-X Knockoffs

We would like to emphasize that the linear model assumption is no longer needed here and arbitrary dependence structure of response $\mathbf{y}$ on covariates $\mathbf{x}$ is allowed. As mentioned in Section 1, to overcome the difficulty caused by unknown precision matrix $\boldsymbol{\Omega}_0$ we modify the model-X knockoffs procedure described in Section 2.1 and suggest the method of graphical nonlinear knockoffs (RANK).

To ease the presentation, we first introduce some notation. For each given $p \times p$ symmetric positive definite matrix $\boldsymbol{\Omega}$, denote by $\mathbf{C}^{\boldsymbol{\Omega}} = \mathbf{I}_p - \text{diag}\{\mathbf{s}\}\boldsymbol{\Omega}$ and $\mathbf{B}^{\boldsymbol{\Omega}} = \left(2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Omega}\text{diag}\{\mathbf{s}\}\right)^{1/2}$ the square root matrix. We define $n \times p$ matrix $\widetilde{\mathbf{X}}^{\boldsymbol{\Omega}} = (\widetilde{\mathbf{x}}_1^{\boldsymbol{\Omega}}, \ldots, \widetilde{\mathbf{x}}_n^{\boldsymbol{\Omega}})^T$ by independently generating $\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}}$ from the conditional distribution

$$\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}}|\mathbf{x}_i \sim N\left(\mathbf{C}^{\boldsymbol{\Omega}}\mathbf{x}_i, (\mathbf{B}^{\boldsymbol{\Omega}})^2\right), \quad (12)$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the original $n \times p$ design matrix generated from Gaussian graphical model (2). It is easy to show that the $(2p)$-variate random vectors $(\mathbf{x}_i^T, (\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}})^T)^T$ are iid with Gaussian distribution of mean $\mathbf{0}$ and covariance matrix given by $\text{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}_0$, $\text{cov}(\mathbf{x}_i, \widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}}) = \boldsymbol{\Sigma}_0\mathbf{C}^{\boldsymbol{\Omega}}$, and $\text{cov}(\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}}) = (\mathbf{B}^{\boldsymbol{\Omega}})^2 + \mathbf{C}^{\boldsymbol{\Omega}}\boldsymbol{\Sigma}_0(\mathbf{C}^{\boldsymbol{\Omega}})^T$.

Our modified knockoffs method RANK exploits the idea of data splitting, in which one half of the sample is used to estimate unknown precision matrix $\boldsymbol{\Omega}_0$ and reduce the model dimensionality, and the other half of the sample is employed to construct the knockoff variables and implement the knockoffs inference procedure, with the steps detailed below.

- *Step 1.* Randomly split the data $(\mathbf{X}, \mathbf{y})$ into 2-folds $(\mathbf{X}^{(k)}, \mathbf{y}^{(k)})$ with $1 \leq k \leq 2$ each of sample size $n/2$.
- *Step 2.* Use the first fold of data $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ to obtain an estimate $\widehat{\boldsymbol{\Omega}}$ of the precision matrix and a reduced model with support $\widetilde{\mathcal{S}}$.
- *Step 3.* With estimated precision matrix $\widehat{\boldsymbol{\Omega}}$ from Step 2, construct an $(n/2) \times p$ knockoffs matrix $\widehat{\mathbf{X}}$ using $\mathbf{X}^{(2)}$ with rows independently generated from Equation (12); that is, $\widehat{\mathbf{X}} = \mathbf{X}^{(2)}(\mathbf{C}^{\widehat{\boldsymbol{\Omega}}})^T + \mathbf{Z}\mathbf{B}^{\widehat{\boldsymbol{\Omega}}}$ with $\mathbf{Z}$ an $(n/2) \times p$ matrix with iid $N(0,1)$ components.
- *Step 4.* Construct knockoff statistics W_j's using only data on support $\widetilde{\mathcal{S}}$, that is, $W_j = W_j(\mathbf{y}^{(2)}, \mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}})$ for $j \in \widetilde{\mathcal{S}}$ and $W_j = 0$ for $j \in \widetilde{\mathcal{S}}^c$. Then apply knockoffs inference procedure to W_j's to obtain final set of features $\widehat{\mathcal{S}}$.

Here for any matrix $\mathbf{A}$ and subset $\mathcal{S} \subset \{1, \ldots, p\}$, the compact notation $\mathbf{A}_{\mathcal{S}}$ stands for the submatrix of $\mathbf{A}$ consisting of columns in set $\mathcal{S}$.

As discussed in Section 2.1, the model-X knockoffs framework uses sparse regression procedures such as the Lasso. For this reason, even in the original model-X knockoffs procedure the knockoff statistics W_j's (see, e.g., Equation (9)) take nonzero values only over a much smaller model than the full model. This observation motivates us to estimate such a smaller model using the first half of the sample in Step 2 of our modified procedure. When implementing this modified procedure, we limit ourselves to sparse models $\widetilde{\mathcal{S}}$ with size bounded by some positive integer K_n that diverges with n; see, for example, Fan and Lv

(2013) and Lv (2013) for detailed discussions and justifications on similar consideration of sparse models. In addition to sparse regression procedures, feature screening methods such as Fan and Lv (2008) and Fan and Fan (2008) can also be used to obtain the reduced model $\widetilde{\mathcal{S}}$.

The above modified knockoffs method differs from the original model-X knockoffs procedure Candès et al. (2018) in that we use an independent sample to obtain the estimated precision matrix $\widehat{\boldsymbol{\Omega}}$ and reduced model $\widetilde{\mathcal{S}}$. In particular, the independence between estimates $(\widehat{\boldsymbol{\Omega}}, \widetilde{\mathcal{S}})$ and data $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ plays an important role in our theoretical analysis for the robustness of the knockoffs procedure. In fact, the idea of data splitting has been popularly used in the literature for various purposes Fan, Samworth, and Wu (2009), Fan, Guo, and Hao (2012), Shah and Samworth (2013), and Barber and Candès (2016). Although the work of Barber and Candès (2016) has the closest connection to ours, there are several key differences between these two methods. Specifically, Barber and Candès (2016) considered high-dimensional linear model with fixed design, where the data is split into two portions with the first portion used for feature screening and the second portion employed for applying the original knockoff filter in Barber and Candès (2015) on the reduced model. To ensure FDR control, it was required in Barber and Candès (2016) that the feature screening method should enjoy the sure screening property Fan and Lv (2008), that is, the reduced model after the screening step contains the true model $\mathcal{S}_0$ with asymptotic probability one. In contrast, one major advantage of our method is that the asymptotic FDR control can be achieved without requiring the sure screening property; see Theorem 2 in Section 3.2 for more details. Such major distinction is rooted on the difference in constructing knockoff variables; that is, we construct model-X knockoff variables *globally* in Step 3 above, whereas Barber and Candès (2016) constructed knockoff variables *locally* on the reduced model. Another major difference is that our method works with random design and does not need any assumption on how response $\mathbf{y}$ depends upon covariates $\mathbf{x}$, while the method in Barber and Candès (2016) requires the linear model assumption and cannot be extended to nonlinear models.

3.2. Robustness of FDR Control for Graphical Nonlinear Knockoffs

We begin with investigating the robustness of FDR control for the modified model-X knockoffs procedure RANK. To simplify the notation, we rewrite $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ as $(\mathbf{X}, \mathbf{y})$ with sample size n whenever there is no confusion, where n now represents half of the original sample size. For each given $p \times p$ symmetric positive definite matrix $\boldsymbol{\Omega}$, an $n \times p$ knockoffs matrix $\widetilde{\mathbf{X}}^{\boldsymbol{\Omega}} = (\widetilde{\mathbf{x}}_1^{\boldsymbol{\Omega}}, \ldots, \widetilde{\mathbf{x}}_n^{\boldsymbol{\Omega}})^T$ can be constructed with n rows independently generated according to Equation (12) and the modified knockoffs procedure proceeds with a given reduced model $\mathcal{S}$. Then the FDP and FDR functions in Equation (1) can be rewritten as

$$\mathrm{FDR}_n(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[\mathrm{FDP}_n(\mathbf{y}, \mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}})], \qquad (13)$$

where the subscript n is used to emphasize the dependence of FDP and FDR functions on sample size. It is easy to check that the knockoffs procedure based on $(\mathbf{y}, \mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}_0})$ satisfies all

the conditions in Candès et al. (2018) for FDR control for any reduced model $\mathcal{S}$, that is, independent of $\mathbf{X}$ and $\widetilde{\mathbf{X}}^{\boldsymbol{\Omega}_0}$, which ensures that $\mathrm{FDR}_n(\boldsymbol{\Omega}_0, \mathcal{S})$ can be controlled at the target level q. To study the robustness of our modified knockoffs procedure, we will make a connection between functions $\mathrm{FDR}_n(\boldsymbol{\Omega}, \mathcal{S})$ and $\mathrm{FDR}_n(\boldsymbol{\Omega}_0, \mathcal{S})$.

To ease the presentation, denote by $\widetilde{\mathbf{X}}_0 = \widetilde{\mathbf{X}}^{\boldsymbol{\Omega}_0}$ the oracle knockoffs matrix with $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0$, $\mathbf{C}_0 = \mathbf{C}^{\boldsymbol{\Omega}_0}$, and $\mathbf{B}_0 = \mathbf{B}^{\boldsymbol{\Omega}_0}$. The following proposition establishes a formal characterization of the FDR as a function of the precision matrix $\boldsymbol{\Omega}$ used in generating the knockoff variables and the reduced model $\mathcal{S}$.

Proposition 1. For any given symmetric positive definite $\boldsymbol{\Omega} \in \mathbb{R}^{p \times p}$ and $\mathcal{S} \subset \{1, \ldots, p\}$, it holds that

$$\mathrm{FDR}_n(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}\left[g_n\left(\mathbf{X}_{\mathrm{aug}}^{\mathcal{S}} \mathbf{H}^{\boldsymbol{\Omega}}\right)\right], \qquad (14)$$

where $\mathbf{X}_{\mathrm{aug}}^{\mathcal{S}} = [\mathbf{X}, \widetilde{\mathbf{X}}_{0,\mathcal{S}}] \in \mathbb{R}^{n \times (p+|\mathcal{S}|)}$, function $g_n(\cdot)$ is some conditional expectation of the FDP function whose functional form is free of $\boldsymbol{\Omega}$ and $\mathcal{S}$, and

$$\mathbf{H}^{\boldsymbol{\Omega}} = \begin{pmatrix} \mathbf{I}_p & \mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}(\mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})^{-1/2} \big((\mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}})^T \mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}}\big)^{1/2} \\ \mathbf{0} & (\mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})^{-1/2}\big((\mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}})^T \mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}}\big)^{1/2} \end{pmatrix}.$$

We see from Proposition 1 that when $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0$, it holds that $\mathbf{H}^{\boldsymbol{\Omega}_0} = \mathbf{I}_{p+|\mathcal{S}|}$ and thus the value of the FDR function at point $\boldsymbol{\Omega}_0$ reduces to

$$\mathrm{FDR}_n(\boldsymbol{\Omega}_0, \mathcal{S}) = \mathbb{E}\left[g_n\left(\mathbf{X}_{\mathrm{aug}}^{\mathcal{S}}\right)\right],$$

which can be shown to be bounded from above by the target FDR level q using the results proved in Candès et al. (2018). Since the dependence of FDR function on $\boldsymbol{\Omega}$ is completely through matrix $\mathbf{H}^{\boldsymbol{\Omega}}$, we can reparameterize the FDR function as $\mathrm{FDR}_n(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S})$. In view of Equation (14), $\mathrm{FDR}_n(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S})$ is the expectation of some measurable function with respect to the probability law of $\mathbf{X}_{\mathrm{aug}}^{\mathcal{S}}$ which has matrix normal distribution with independent rows, and thus is expected to be a smooth function of entries of $\mathbf{H}^{\boldsymbol{\Omega}}$ by measure theory. Motivated by such an observation, we make the following Lipschitz continuity assumption.

Condition 4. There exists some constant $L > 0$ such that for all $|\mathcal{S}| \le K_n$ and $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 \le C_2 a_n$ with some constant $C_2 > 0$ and $a_n \to 0$, $\big|\mathrm{FDR}_n(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S}) - \mathrm{FDR}_n(\mathbf{H}^{\boldsymbol{\Omega}_0}, \mathcal{S})\big| \le L\|\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0}\|_F$, where $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the matrix spectral norm and matrix Frobenius norm, respectively.

Condition 5. Assume that the estimated precision matrix $\widehat{\boldsymbol{\Omega}}$ satisfies $\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_2 \le C_2 a_n$ with probability $1 - O(p^{-c_1})$ for some constants $C_2, c_1 > 0$ and $a_n \to 0$, and that $|\widehat{\mathcal{S}}| \le K_n$.

The error rate of precision matrix estimation assumed in Condition 5 is quite flexible. We would like to emphasize that no sparsity assumption has been made on the true precision matrix $\boldsymbol{\Omega}_0$. Bounding the size of sparse models is also important for ensuring model identifiability and stability; see, for instance, Fan and Lv (2013) and Lv (2013) for more detailed discussions.

Theorem 2. Assume that all the eigenvalues of $\mathbf{\Omega}_0$ are bounded away from 0 to ∞ and the smallest eigenvalue of $2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\mathbf{\Omega}_0\text{diag}(\mathbf{s})$ is bounded from below by some positive constant. Then under Condition 4, it holds that

$$\sup_{|\mathcal{S}| \leq K_n, \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_2 \leq C_2 a_n} |\text{FDR}_n(\mathbf{H}^{\mathbf{\Omega}}, \mathcal{S}) - \text{FDR}(\mathbf{H}^{\mathbf{\Omega}_0}, \mathcal{S})|$$
$$\leq O(K_n^{1/2} a_n). \tag{15}$$

Moreover, under Conditions 4 and 5 with $K_n^{1/2} a_n \to 0$, the FDR of RANK is bounded from above by $q + O(K_n^{1/2} a_n) + O(p^{-c_1})$, where $q \in (0, 1)$ is the target FDR level.

Theorem 2 establishes the robustness of the FDR with respect to the precision matrix $\mathbf{\Omega}$; see the uniform bound in Equation (15). As a consequence, it shows that our modified model-X knockoffs procedure RANK can indeed have FDR asymptotically controlled at the target level q. We remark that the term $K_n^{1/2}$ in Theorem 2 is because Condition 4 is imposed through the matrix Frobenius norm, which is motivated from results on the smoothness of integral function from calculus. If one is willing to impose assumption through matrix spectral norm instead of Frobenius norm, then the extra term $K_n^{1/2}$ can be dropped and the set $\mathcal{S}$ can be taken as the full model $\{1, \ldots, p\}$.

We would like to stress that Theorem 2 allows for arbitrarily complicated dependence structure of response $\mathbf{y}$ on covariates $\mathbf{x}$ and for any valid construction of knockoff statistics W_j's. This is different from the conditions needed for power analysis in Section 2.2 (i.e., the linear model setting and LCD knockoff statistics). Moreover, the asymptotic FDR control in Theorem 2 does not need the sure screening property of $\mathbb{P}\{\widetilde{\mathcal{S}} \supset \mathcal{S}_0\} \to 1$ as $n \to \infty$.

3.3. Robustness of Power in Linear Models

We are now curious about the other side of the coin; that is, the robustness theory for the power of our modified knockoffs procedure RANK. As argued at the beginning of Section 2.2, to ease the presentation and simplify the technical derivations we come back to high-dimensional linear models (8) and use the LCD in Equation (9) as the knockoff statistics. The difference with the setting in Section 2.2 is that we no longer assume that the true precision matrix $\mathbf{\Omega}_0$ is known and use the modified knockoffs procedure introduced in Section 3.1 to achieve asymptotic FDR control.

Recall that for the RANK procedure, the reduced model $\widetilde{\mathcal{S}}$ is first obtained from an independent subsample and then the knockoffs procedure is applied on the second fold of data to further select features from $\widetilde{\mathcal{S}}$. Clearly, if $\widetilde{\mathcal{S}}$ does not have the sure screening property of $\mathbb{P}\{\widetilde{\mathcal{S}} \supset \mathcal{S}_0\} \to 1$ as $n \to \infty$, then the Lasso solution based on $[\mathbf{X}^{(2)}_{\widetilde{\mathcal{S}}}, \widetilde{\mathbf{X}}^{\mathbf{\Omega}}_{\widetilde{\mathcal{S}}}]$ as given in Equation (18) is no longer a consistent estimate of $\boldsymbol{\beta}_0$ even when the true precision matrix $\mathbf{\Omega}_0$ is used to generate the knockoff variables. In addition, the final power of our modified knockoffs procedure will always be upper bounded by $s^{-1}|\widetilde{\mathcal{S}} \cap \mathcal{S}_0|$. Nevertheless, the results in this section are still useful in the sense that model (8) can be viewed as the projected model on support $\widetilde{\mathcal{S}}$. Thus, our power analysis here is relative power analysis with respect to the reduced model $\widetilde{\mathcal{S}}$. In other words, we will focus on how much

power loss would occur after we apply the model-X knockoffs procedure to $(\mathbf{X}^{(2)}_{\widetilde{\mathcal{S}}}, \widetilde{\mathbf{X}}^{\mathbf{\Omega}}_{\widetilde{\mathcal{S}}}, \mathbf{y}^{(2)})$ when compared to the power of $s^{-1}|\widetilde{\mathcal{S}} \cap \mathcal{S}_0|$. Since our focus is relative power loss, without loss of generality we will condition on the event

$$\{\widetilde{\mathcal{S}} \supset \mathcal{S}_0\}. \tag{16}$$

We would like to point out that all conditions and results in this section can be adapted correspondingly when we view model (8) as the projected model if $\widetilde{\mathcal{S}} \not\supset \mathcal{S}_0$. Similarly as in FDR analysis, we restrict ourselves to sparse models with size bounded by K_n that diverges as $n \to \infty$, that is, $|\widetilde{\mathcal{S}}| \leq K_n$.

With $\mathbf{\Omega}$ taken as the estimated precision matrix $\widehat{\mathbf{\Omega}}$, we can generate the knockoff variables from Equation (12). Then the Lasso procedure can be applied to the augmented data $(\mathbf{X}^{(2)}, \widehat{\mathbf{X}}, \mathbf{y}^{(2)})$ with $\widehat{\mathbf{X}}$ constructed in Step 3 of our modified knockoffs procedure and the LCD can be defined as

$$\widehat{W}_j = W_j^{\widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}}} = |\widehat{\beta}_j(\lambda; \widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}})| - |\widehat{\beta}_{p+j}(\lambda; \widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}})|, \tag{17}$$

where $\widehat{\beta}_j(\lambda; \widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}})$ and $\widehat{\beta}_{p+j}(\lambda; \widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}})$ are the jth and $(j + p)$th components, respectively, of the Lasso estimator

$$\widehat{\boldsymbol{\beta}}(\lambda; \widehat{\mathbf{\Omega}}, \widetilde{\mathcal{S}}) = \text{argmin}_{\mathbf{b}_{\widetilde{\mathcal{S}}_1} = \mathbf{0}} \left\{ n^{-1} \|\mathbf{y}^{(2)} - [\mathbf{X}^{(2)}, \widehat{\mathbf{X}}]\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\} \tag{18}$$

with $\lambda \geq 0$ the regularization parameter and $\widetilde{\mathcal{S}}_1 = \{1 \leq j \leq 2p : j \notin \widetilde{\mathcal{S}} \text{ and } j - p \notin \widetilde{\mathcal{S}}\}$.

Unlike the FDR analysis in Section 3.2, we now need sparsity assumption on the true precision matrix $\mathbf{\Omega}_0$.

Condition 6. Assume that $\mathbf{\Omega}_0$ is L_p-sparse with each row having at most L_p nonzeros for some diverging L_p and all the eigenvalues of $\mathbf{\Omega}_0$ are bounded away from 0 to ∞.

For each given precision matrix $\mathbf{\Omega}$ and reduced model $\mathcal{S}$, we define $W_j^{\mathbf{\Omega}, \mathcal{S}}$ similarly as in (17) except that $\mathbf{\Omega}$ is used to generate the knockoff variables and set $\mathcal{S}$ is used in Equation (18) to calculate the Lasso solution. Denote by $\widehat{\mathcal{S}}^{\mathbf{\Omega}} = \{j : W_j^{\mathbf{\Omega}, \mathcal{S}} \geq T\} \subset \mathcal{S}$ the final set of selected features using the LCD $W_j^{\mathbf{\Omega}, \mathcal{S}}$ in the knockoffs inference framework. We further define a class of precision matrices $\mathbf{\Omega} \in \mathbb{R}^{p \times p}$

$$\mathcal{A} = \left\{ \mathbf{\Omega} : \mathbf{\Omega} \text{ is } L_p'\text{-sparse and } \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_2 \leq C_2 a_n \right\}, \tag{19}$$

where C_2 and a_n are the same as in Theorem 2 and L_p' is some positive integer that diverges with n. Similarly as in Section 2.2, in the technical analysis we assume implicitly that with asymptotic probability one, for all valid constructions of the knockoff variables there are no ties in the magnitude of nonzero knockoff statistics and no ties in the magnitude of nonzero components of Lasso solution uniformly over all $\mathbf{\Omega} \in \mathcal{A}$ and $|\mathcal{S}| \leq K_n$.

Condition 7. It holds that $\mathbb{P}\{\widehat{\mathbf{\Omega}} \in \mathcal{A}\} \geq 1 - c_2 p^{-c_2}$ for some constant $c_2 > 0$.

The assumption on the estimated precision matrix $\widehat{\mathbf{\Omega}}$ made in Condition 7 is mild and flexible. A similar class of precision matrices was considered in Fan et al. (2015) with detailed discussions on the choices of the estimation procedures. See, for

example, Fan and Lv (2016), Chen et al. (2016), and Ren et al. (2018) for some more recent developments on large precision matrix estimation and inference. In parallel to Theorem 1, we have the following results on the power of our modified knockoffs procedure with the estimated precision matrix $\widehat{\boldsymbol{\Omega}}$.

Theorem 3. Assume that Conditions 1 and 2, and 6 and 7 hold, the smallest eigenvalue of $2\text{diag}(\mathbf{s})-\text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})$ is positive and bounded away from 0, $|\{j : |\beta_{0,j}| \gg [sn^{-1}(\log p)]^{1/2}\}| \geq cs$, and $\lambda = C_\lambda\{(\log p)/n\}^{1/2}$ with $c \in ((qs)^{-1}, 1)$ and $C_\lambda > 0$ some constants. Then if $[(L_p + L'_p)^{1/2} + K_n^{1/2}]a_n = o(1)$ and $s\{a_n + (K_n + L'_p)[n^{-1}(\log p)]^{1/2}\} = o(1)$, RANK with estimated precision matrix $\widehat{\boldsymbol{\Omega}}$ and reduced model $\widetilde{\mathcal{S}}$ has power satisfying

$$\text{Power}(\widehat{\boldsymbol{\Omega}}, \widetilde{\mathcal{S}}) = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}}^{\widehat{\boldsymbol{\Omega}}} \cap \mathcal{S}_0|}{|\mathcal{S}_0|}\right] \geq 1 - C_{\ell_1}C_\lambda(\varphi + 1)\kappa_n^{-1}$$
$$- c_2 p^{-c_2} - c_3 p^{-c_3} + o(\kappa_n^{-1})$$
$$= 1 - o(1),$$

where φ is the golden ratio and C_{ℓ_1} is some positive constant.

Theorem 3 establishes the robustness of the power for the RANK method. In view of Theorems 2 and 3, we see that our modified knockoffs procedure RANK can enjoy appealing properties of FDR control and power simultaneously when the true covariate distribution is unknown and needs to be estimated in high dimensions.

4. Simulation Studies

So far we have seen that our suggested RANK method admits appealing theoretical properties for large-scale inference in high-dimensional nonlinear models. We now examine the finite-sample performance of RANK through four simulation examples.

4.1. Model Setups and Simulation Settings

Recall that the original knockoff filter (KF) in Barber and Candès (2015) was designed for linear regression model with dimensionality p not exceeding sample size n, while the high-dimensional knockoff filter (HKF) in Barber and Candès (2016) considers linear model with p possibly larger than n. To compare RANK with the HKF procedure in high-dimensional setting, our first simulation example adopts the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{20}$$

where $\mathbf{y}$ is an n-dimensional response vector, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional regression coefficient vector, and $\boldsymbol{\varepsilon}$ is an n-dimensional error vector. Nonlinear models provide useful and flexible alternatives to linear models and are widely used in real applications. Our second through fourth simulation examples are devoted to three popular nonlinear model settings: the partially linear model, the single-index model, and the additive model, respectively. As a natural extension of linear model (20), the partially linear model assumes that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g}(\mathbf{U}) + \boldsymbol{\varepsilon}, \tag{21}$$

where $\mathbf{g}(\mathbf{U}) = (g(U_1), \ldots, g(U_n))^T$ is an n-dimensional vector-valued function with covariate vector $\mathbf{U} = (U_1, \ldots, U_n)^T, g(\cdot)$ is some *unknown* smooth nonparametric function, and the rest of notation is the same as in model (20). In particular, the partially linear model is a semiparametric regression model that has been commonly used in many areas such as economics, finance, medicine, epidemiology, and environmental science Engle et al. (1986) and Härdle, Liang, and Gao (2000).

The third and fourth simulation examples drop the linear component. As a popular tool for dimension reduction, the single-index model assumes that

$$\mathbf{y} = \mathbf{g}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}, \tag{22}$$

where $\mathbf{g}(\mathbf{X}\boldsymbol{\beta}) = (g(\mathbf{x}_1^T\boldsymbol{\beta}), \ldots, g(\mathbf{x}_n^T\boldsymbol{\beta}))^T$ with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, $g(\cdot)$ is an *unknown* link function, and the remaining notation is the same as in model (20). In particular, the single-index model provides a flexible extension of the GLM by relaxing the parametric form of the link function Ichimura (1993), Stoker (1986), Härdle and Stoker (1989), Li and Racine (2007), and Horowitz (2009). To bring more flexibility while alleviating the curse of dimensionality, the additive model assumes that

$$\mathbf{y} = \sum_{j=1}^{p} \mathbf{g}_j(\mathbf{X}_j) + \boldsymbol{\varepsilon}, \tag{23}$$

where $\mathbf{g}_j(\boldsymbol{\theta}) = (g_j(\theta_1), \ldots, g_j(\theta_n))^T$ for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, $\mathbf{X}_j$ represents the jth covariate vector with $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, $g_j(\cdot)$'s are some *unknown* smooth functions, and the rest of notation is the same as in model (20). The additive model has been widely employed for nonparametric modeling of high-dimensional data Hastie and Tibshirani (1990), Ravikumar et al. (2009), Meier, van de Geer, and Bühlmann (2009), and Chouldechova and Hastie (2015).

For the linear model (20) in simulation example 1, the rows of the $n \times p$ design matrix $\mathbf{X}$ are generated as iid copies of $N(\mathbf{0}, \boldsymbol{\Sigma})$ with precision matrix $\boldsymbol{\Sigma}^{-1} = (\rho^{|j-k|})_{1 \leq j,k \leq p}$ for $\rho = 0$ and 0.5. We set the true regression coefficient vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ as a sparse vector with $s = 30$ nonzero components, where the signal locations are chosen randomly and each nonzero coefficient is selected randomly from $\{\pm A\}$ with $A = 1.5$ and 3.5. The error vector $\boldsymbol{\varepsilon}$ is assumed to be $N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ with $\sigma = 1$. We set sample size $n = 400$ and consider the high-dimensional scenario with dimensionality $p = 200, 400, 600, 800$, and 1000. For the partially linear model (21) in simulation example 2, we choose the true function as $g(U) = \sin(2\pi U)$, generate $\mathbf{U} = (U_1, \ldots, U_n)^T$ with iid U_i from uniform distribution on $[0, 1]$, and set $A = 1.5$ with the remaining setting the same as in simulation example 1.

Since the single-index model and additive model are more complex than the linear model and partially linear model, we reduce the true model size s while keeping sample size $n = 400$ in both simulation examples 3 and 4. For the single-index model (22) in simulation example 3, we consider the true link function $g(x) = x^3/2$ and set $p = 200, 400, 600, 800$, and 1000. The true p-dimensional regression coefficient vector $\boldsymbol{\beta}_0$ is generated similarly with $s = 10$ and $A = 1.5$. For the additive model (23) in simulation example 4, we assume that $s = 10$ of the functions $g_j(\cdot)$'s are nonzero with j's chosen randomly from $\{1, \ldots, p\}$ and

the remaining $p - 10$ functions $g_j(\cdot)$'s vanish. Specifically, each nonzero function $g_j(\cdot)$ is taken to be a polynomial of degree 3 and all coefficients under the polynomial basis functions are generated independently as $N(0, 10^2)$ as in Chouldechova and Hastie (2015). The dimensionality p is allowed to vary with values $200, 400, 600, 800,$ and 1000. For each simulation example, we set the number of repetitions as 100.

4.2. Estimation Procedures

To implement RANK procedure described in Section 3.1, we need to construct a precision matrix estimator $\widehat{\boldsymbol{\Omega}}$ and obtain the reduced model $\widetilde{\mathcal{S}}$ using the first fold of data $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$. Among all available estimators in the literature, we employ the ISEE method in Fan and Lv (2016) for precision matrix estimation due to its scalability, simple tuning, and nice theoretical properties. For simplicity, we choose $s_j = 1/\Lambda_{\max}(\widehat{\boldsymbol{\Omega}})$ for all $1 \leq j \leq p$, where $\widehat{\boldsymbol{\Omega}}$ denotes the ISEE estimator for the true precision matrix $\boldsymbol{\Omega}_0$ and $\Lambda_{\max}$ standards for the largest eigenvalue of a matrix. Then we can obtain an $(n/2) \times (2p)$ augmented design matrix $[\mathbf{X}^{(2)}, \widehat{\mathbf{X}}]$, where $\widehat{\mathbf{X}}$ represents an $(n/2) \times p$ knockoffs matrix constructed in Step 3 of our modified knockoffs procedure in Section 3.1. To construct the reduced model $\widetilde{\mathcal{S}}$ using the first fold of data $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$, we borrow the strengths from the recent literature on feature selection methods. After $\widetilde{\mathcal{S}}$ is obtained, we employ the reduced data $(\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}}, \mathbf{y}^{(2)})$ with $\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}} = [\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}}]$ to fit a model and construct the knockoff statistics. In what follows, we will discuss feature selection methods for obtaining $\widetilde{\mathcal{S}}$ for the linear model (20), partially linear model (21), single-index model (22), and additive model (23) in simulation examples 1–4, respectively. We will also discuss the construction of knockoff statistics in each model setting.

For the linear model (20) in simulation example 1, we obtain the reduced model $\widetilde{\mathcal{S}}$ by first applying the Lasso procedure

$$\widehat{\boldsymbol{\beta}}^{(1)} = \text{argmin}_{\mathbf{b} \in \mathbb{R}^p} \left\{ n^{-1} \|\mathbf{y}^{(1)} - \mathbf{X}^{(1)} \mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\} \quad (24)$$

with $\lambda \geq 0$ the regularization parameter and then taking the support $\widetilde{\mathcal{S}} = \text{supp}(\widehat{\boldsymbol{\beta}}^{(1)})$. Then with the estimated $\widehat{\boldsymbol{\Omega}}$ and $\widetilde{\mathcal{S}}$, we construct the knockoff statistics as the LCD (17), where the estimated regression coefficient vector is obtained by applying the Lasso procedure on the reduced model as described in (18). The regularization parameter λ in Lasso is tuned using the K-fold cross-validation (CV).

For the partially linear model (21) in simulation example 2, we employ the profiling method in semiparametric regression based on the first fold of data $(\mathbf{X}^{(1)}, \mathbf{U}^{(1)}, \mathbf{y}^{(1)})$ by observing that model (21) becomes a linear model when conditioning on the covariate vector $\mathbf{U}^{(1)}$. Consequently, we need to estimate both the profiled response $\mathbb{E}(\mathbf{y}^{(1)} | \mathbf{U}^{(1)})$ and the profiled covariates $\mathbb{E}(\mathbf{X}^{(1)} | \mathbf{U}^{(1)})$. To this end, we adopt the local linear smoothing estimators Fan and Gijbels (1996) $\widehat{\mathbb{E}(\mathbf{y}^{(1)} | \mathbf{U}^{(1)})}$ and $\widehat{\mathbb{E}(\mathbf{X}^{(1)} | \mathbf{U}^{(1)})}$ of $\mathbb{E}(\mathbf{y}^{(1)} | \mathbf{U}^{(1)})$ and $\mathbb{E}(\mathbf{X}^{(1)} | \mathbf{U}^{(1)})$ using the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ with the optimal bandwidth selected by the generalized cross-validation (GCV). Then we define the Lasso estimator $\widehat{\boldsymbol{\beta}}^{(1)}$ for the p-dimensional regression coefficient vector similarly as in Equation (24) with $\mathbf{y}^{(1)}$ and $\mathbf{X}^{(1)}$ replaced by $\mathbf{y}^{(1)} - \widehat{\mathbb{E}(\mathbf{y}^{(1)} | \mathbf{U}^{(1)})}$ and $\mathbf{X}^{(1)} - \widehat{\mathbb{E}(\mathbf{X}^{(1)} | \mathbf{U}^{(1)})}$, respectively. The

reduced model is then taken as $\widetilde{\mathcal{S}} = \text{supp}(\widehat{\boldsymbol{\beta}}^{(1)})$. For knockoff statistics $\widehat{W}_j$, we set $\widehat{W}_j = 0$ for all $j \notin \widetilde{\mathcal{S}}$. On the support $\widetilde{\mathcal{S}}$, we construct $\widehat{W}_j = |\widehat{\beta}_j| - |\widehat{\beta}_{p+j}|$ with $\widehat{\beta}_j$ and $\widehat{\beta}_{p+j}$ the Lasso coefficients obtained by applying the model fitting procedure described above to the reduced data $(\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}}, \mathbf{U}_{\widetilde{\mathcal{S}}}^{(2)}, \mathbf{y}^{(2)})$ in the second subsample with $\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}} = [\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}}]$.

To fit the single-index model (22) in simulation example 3, we employ the Lasso-SIR method in Lin, Zhao, and Liu (2016). The Lasso-SIR first divides the sample of $m = n/2$ observations in the first subsample $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ into H slices of equal length c, and constructs the matrix $\Lambda_H = \frac{1}{mc}(\mathbf{X}^{(1)})^T \mathbf{M} \mathbf{M}^T \mathbf{X}^{(1)}$, where $\mathbf{M} = \mathbf{I}_H \otimes \mathbf{1}_c$ is an $m \times H$ matrix, that is, the Kronecker product of the identity matrix $\mathbf{I}_H$ and the constant vector $\mathbf{1}_c$ of ones. Then the Lasso-SIR estimates the p-dimensional regression coefficient vector $\widehat{\boldsymbol{\beta}}^{(1)}$ using the Lasso procedure similarly as in Equation (18) with the original response vector $\mathbf{y}^{(1)}$ replaced by a new response vector $\widetilde{\mathbf{y}}^{(1)} = (c\lambda_1)^{-1} \mathbf{M} \mathbf{M}^T \mathbf{X}^{(1)} \boldsymbol{\eta}_1$, where λ_1 denotes the largest eigenvalue of matrix Λ_H and $\boldsymbol{\eta}_1$ is the corresponding eigenvector. We set the number of slices $H = 5$. Then the reduced model is taken as $\widetilde{\mathcal{S}} = \text{supp}(\widehat{\boldsymbol{\beta}}^{(1)})$. We then apply the fitting procedure Lasso-SIR discussed above to the reduced data $(\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}}, \mathbf{y}^{(2)})$ with $\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}} = [\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}}]$ and construct knockoff statistics in a similar way as in partially linear model.

To fit the additive model (23) in simulation example 4, we apply the GAMSEL procedure in Chouldechova and Hastie (2015) for sparse additive regression. In particular, we choose 6 basis functions each with 6 degrees of freedom for the smoothing splines using orthogonal polynomials for each additive component and set the penalty mixing parameter $\gamma = 0.9$ in GAMSEL to obtain estimators of the true functions $g_j(\cdot)$'s. The GAMSEL procedure is first applied to the first subsample $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$ to obtain the reduced model $\widetilde{\mathcal{S}}$, and then applied to the reduced data $(\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}}, \mathbf{y}^{(2)})$ with $\mathbf{X}_{\text{aug}}^{\widetilde{\mathcal{S}}} = [\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}}]$ to obtain estimates $\widehat{g}_j$ and $\widehat{g}_{p+j}$ for the additive functions corresponding to the jth covariate and its knockoff counterpart with $j \in \widetilde{\mathcal{S}}$, respectively. The knockoff statistics are then constructed as

$$\widehat{W}_j = \|\widehat{g}_j\|_{n/2}^2 - \|\widehat{g}_{P+j}\|_{n/2}^2 \text{ for } j \in \widetilde{\mathcal{S}} \quad (25)$$

and $\widehat{W}_j = 0$ for $j \notin \widetilde{\mathcal{S}}$, where $\|\widehat{g}_j\|_{n/2}$ represents the empirical norm of the estimated function $\widehat{g}_j(\cdot)$ evaluated at its observed points and $n/2$ stands for the size of the second subsample.

It is seen that in all four examples above, intuitively large positive values of knockoff statistics $\widehat{W}_j$ provide strong evidence against the jth null hypothesis $H_{0,j} : \beta_j = 0$ or $H_{0,j} : g_j = 0$. For all simulation examples, we set the target FDR level at $q = 0.2$.

4.3. Simulation Results

To gain some insights into the effect of data splitting, we also implemented our procedure without the data splitting step. To differentiate, we use RANKs to denote the procedure with data splitting and RANK to denote the procedure without data splitting. To examine the feature selection performance, we look at both measures of FDR and power. The empirical versions of FDR and power based on 100 replications are reported in

Table 1. Simulation results for linear model (20) in simulation example 1 with $A = 1.5$ in Section 4.1.

		RANK		RANK$_+$		RANKs		RANKs$_+$	
ρ	p	FDR	Power	FDR	Power	FDR	Power	FDR	Power
0	200	0.2054	1.00	0.1749	1.00	0.1909	1.00	0.1730	1.00
	400	0.2062	1.00	0.1824	1.00	0.2010	1.00	0.1801	1.00
	600	0.2263	1.00	0.1940	1.00	0.2206	1.00	0.1935	1.00
	800	0.2385	1.00	0.1911	1.00	0.2247	1.00	0.1874	1.00
	1000	0.2413	1.00	0.2083	1.00	0.2235	1.00	0.1970	1.00
0.5	200	0.2087	1.00	0.1844	1.00	0.1875	1.00	0.1692	1.00
	400	0.2144	1.00	0.1879	1.00	0.1954	1.00	0.1703	1.00
	600	0.2292	1.00	0.1868	1.00	0.2062	1.00	0.1798	1.00
	800	0.2398	1.00	0.1933	1.00	0.2052	0.9997	0.1805	0.9997
	1000	0.2412	1.00	0.2019	1.00	0.2221	0.9984	0.2034	0.9984

Table 2. Simulation results for linear model (20) in simulation example 1 with $A = 3.5$ in Section 4.1.

		RANKs		RANKs$_+$		HKF		HKF$_+$	
ρ	p	FDR	Power	FDR	Power	FDR	Power	FDR	Power
0	200	0.1858	1.00	0.1785	1.00	0.1977	0.9849	0.1749	0.9837
	400	0.1895	1.00	0.1815	1.00	0.2064	0.9046	0.1876	0.8477
	600	0.2050	1.00	0.1702	1.00	0.1964	0.8424	0.1593	0.7668
	800	0.2149	1.00	0.1921	1.00	0.1703	0.7513	0.1218	0.6241
	1000	0.2180	1.00	0.1934	1.00	0.1422	0.7138	0.1010	0.5550
0.5	200	0.1986	1.00	0.1618	1.00	0.1992	0.9336	0.1801	0.9300
	400	0.1971	1.00	0.1805	1.00	0.1657	0.8398	0.1363	0.7825
	600	0.2021	1.00	0.1757	1.00	0.1253	0.7098	0.0910	0.6068
	800	0.2018	1.00	0.1860	1.00	0.1374	0.6978	0.0917	0.5792
	1000	0.2097	0.9993	0.1920	0.9993	0.1552	0.6486	0.1076	0.5524

Table 3. Simulation results for partially linear model (21) in simulation example 2 in Section 4.1.

		RANK		RANK$_+$		RANKs		RANKs$_+$	
ρ	p	FDR	Power	FDR	Power	FDR	Power	FDR	Power
0	200	0.2117	1.00	0.1923	1.00	0.1846	0.9976	0.1699	0.9970
	400	0.2234	1.00	0.1977	1.00	0.1944	0.9970	0.1747	0.9966
	600	0.2041	1.00	0.1776	1.00	0.2014	0.9968	0.1802	0.9960
	800	0.2298	1.00	0.1810	1.00	0.2085	0.9933	0.1902	0.9930
	1000	0.2322	1.00	0.1979	1.00	0.2113	0.9860	0.1851	0.9840
0.5	200	0.2180	1.00	0.1929	1.00	0.1825	0.9952	0.1660	0.9949
	400	0.2254	1.00	0.1966	1.00	0.1809	0.9950	0.1628	0.9948
	600	0.2062	1.00	0.1814	1.00	0.2038	0.9945	0.1898	0.9945
	800	0.2264	1.00	0.1948	1.00	0.2019	0.9916	0.1703	0.9906
	1000	0.2316	1.00	0.2033	1.00	0.2127	0.9830	0.1857	0.9790

Table 4. Simulation results for single-index model (22) in simulation example 3 in Section 4.1.

		RANK		RANK$_+$		RANKs		RANKs$_+$	
ρ	p	FDR	Power	FDR	Power	FDR	Power	FDR	Power
0	200	0.1893	1	0.1413	1	0.1899	1	0.1383	1
	400	0.2163	1	0.1598	1	0.245	0.998	0.1676	0.997
	600	0.2166	1	0.1358	1	0.2314	0.999	0.1673	0.998
	800	0.1964	1	0.1406	1	0.2443	0.992	0.1817	0.992
	1000	0.2051	1	0.134	1	0.2431	0.969	0.1611	0.962
0.5	200	0.2189	1	0.1591	1	0.2322	1	0.1626	1
	400	0.2005	1	0.1314	1	0.2099	0.996	0.1615	0.995
	600	0.2064	1	0.1426	1	0.2331	0.998	0.1726	0.998
	800	0.2049	1	0.1518	1	0.2288	0.994	0.1701	0.994
	1000	0.2259	1	0.1423	1	0.2392	0.985	0.185	0.983

Tables 1 and 2 for simulation example 1 and Tables 3 and 5 for simulation examples 2–4, respectively. In particular, Table 1 compares the performance of RANK and RANK$_+$ with that of RANKs and RANKs$_+$, where the subscript $+$ stands for the corresponding method when the modified knockoff threshold T_+ is used. We see from Table 1 that RANK and RANK$_+$ mimic closely RANKs and RANKs$_+$, respectively, suggesting that data splitting is more of a technical assumption. In addition, the FDR is approximately controlled at the target level of $q = 0.2$ with high power, which is in line with our theory. Table 2 summarizes the comparison of RANKs with HKF procedure for high-dimensional linear regression model. Despite that both methods are based on data splitting, their practical performance is very different. It is seen that although controlling the FDR below the

Table 5. Simulation results for additive model (23) in simulation example 4 in Section 4.1.

ρ	p	RANK		RANK$_+$		RANKs		RANKs$_+$	
		FDR	Power	FDR	Power	FDR	Power	FDR	Power
0	200	0.1926	0.9780	0.1719	0.9690	0.2207	0.9490	0.1668	0.9410
	400	0.2094	0.9750	0.1773	0.9670	0.2236	0.9430	0.1639	0.9340
	600	0.2155	0.9670	0.1729	0.9500	0.2051	0.9310	0.1620	0.9220
	800	0.2273	0.9590	0.1825	0.9410	0.2341	0.9280	0.1905	0.9200
	1000	0.2390	0.9570	0.1751	0.9350	0.2350	0.9140	0.1833	0.9070
0.5	200	0.1904	0.9680	0.1733	0.9590	0.2078	0.9370	0.1531	0.9330
	400	0.2173	0.9650	0.1701	0.9540	0.2224	0.9360	0.1591	0.9280
	600	0.2267	0.9600	0.1656	0.9360	0.2366	0.9340	0.1981	0.9270
	800	0.2306	0.9540	0.1798	0.9320	0.2332	0.9150	0.1740	0.9110
	1000	0.2378	0.9330	0.1793	0.9270	0.2422	0.8970	0.1813	0.8880

target level, HKF suffers from a loss of power due to the use of the screening step and the power deteriorates as dimensionality p increases. In contrast, the performance of RANKs is robust across different correlation levels ρ and dimensionality p. It is worth mentioning that HKF procedure with data recycling performed generally better than that with data splitting alone. Thus, only the results for the former version are reported in Table 2 for simplicity.

For high-dimensional nonlinear settings of partially linear model, single-index model, and additive model in simulation examples 2–4, we see from Tables 3–5 that RANKs and RANKs$_+$ performed well and similarly as RANK and RANK$_+$ in terms of both FDR control and power across different scenarios. These results demonstrate the model-X feature of our procedure for large-scale inference in nonlinear models.

5. Real Data Analysis

In addition to simulation examples presented in Section 4, we also demonstrate the practical utility of our RANK procedure on a gene expression dataset, which is based on Affymetrix GeneChip microarrays for the plant Arabidopsis thaliana in Wille et al. (2004). It is well known that isoprenoids play a key role in plant and animal physiological processes, such as photosynthesis, respiration, regulation of growth, and defense against pathogens in plant physiological processes. In particular, Horvath, Schaffer, and Wisman (2003) found that many of the genes expressed preferentially in mature leaves are readily recognizable as genes involved in photosynthesis, including rubisco activase (AT2G39730), fructose bisphosphate aldolase (AT4G38970), and two glycine hydroxymethyltransferase genes (AT4G37930 and AT5G26780). Thus, isoprenoids have become important ingredients in various drugs (e.g., against cancer and malaria), fragrances (e.g., menthol), and food colorants (e.g., carotenoids). See, for instance, Wille et al. (2004), Schäfer and Strimmer (2005), and Prelić et al. (2006) on studying the mechanism of isoprenoid synthesis in a wide range of applications.

The aforementioned dataset in Wille et al. (2004) consists of 118 gene expression patterns under various experimental conditions for 39 isoprenoid genes, 15 of which are assigned to the regulatory pathway, 19 to the plastidial pathway, and the remaining 5 isoprenoid genes encode protein located in the mitochondrion. Moreover, 795 additional genes from 56 metabolic pathways are incorporated into the isoprenoid genetic network. Thus, the combined dataset is comprised of a sample of

Table 6. Selected genes and their associated pathways for real data analysis in Section 5.

RANK		RANK$_+$	
Pathway	Gene	Pathway	Gene
Calvin	AT4G38970	Calvin	AT4G38970
Carote	AT1G57770	Carote	AT1G57770
Folate	AT1G78670	Folate	AT1G78670
Inosit	AT3G56960		
Phenyl	AT2G27820	Phenyl	AT2G27820
Purine	AT3G01820	Purine	AT3G01820
Ribofl	AT4G13700		
Ribofl	AT2G01880	Ribofl	AT2G01880
Starch	AT5G19220	Starch	AT5G19220

Lasso			
Pathway	Gene	Pathway	Gene
Berber	AT2G34810	Porphy	AT4G18480
Calvin	AT4G38970	Pyrimi	AT5G59440
Calvin	AT3G04790	Ribofl	AT2G01880
Glutam	AT5G18170	Starch	AT5G19220
Glycol	AT4G27600	Starch	AT2G21590
Pentos	AT3G04790	Trypto	AT5G48220
Phenyl	AT2G27820	Trypto	AT5G17980
Porphy	AT1G03475	Mevalo	AT5G47720
Porphy	AT3G51820		

$n = 118$ gene expression patterns for 834 genes. This dataset was studied in Yang, Lozano, and Ravikumar (2014) for identifying genes that exhibit significant association with the specific isoprenoid gene GGPPS11 (AGI code AT4G36810). Motivated by Yang, Lozano, and Ravikumar (2014), we choose the expression level of isoprenoid gene GGPPS11 as the response and treat the remaining $p = 833$ genes from 58 different metabolic pathways as the covariates, in which the dimensionality p is much larger than sample size n. All the variables are logarithmically transformed. To identify important genes associated with isoprenoid gene GGPPS11, we employ the RANK method using the Lasso procedure with target FDR level $q = 0.2$. The implementation of RANK is the same as that in Section 4 for the linear model. Since the sample size of this dataset is relatively low, we choose to implement RANK without sample splitting, which has been demonstrated in Section 4 to be capable of controlling the FDR at the desired level.

Table 6 lists the selected genes by RANK, RANK$_+$, and Lasso along with their associated pathways. We see from Table 6 that RANK, RANK$_+$, and Lasso selected 9 genes, 7 genes, and 17 genes, respectively. The common set of four genes, AT4G38970, AT2G27820, AT2G01880, and AT5G19220, was

selected by all three methods. The values of the adjusted R^2 for these three selected models are equal to 0.7523, 0.7515, and 0.7843, respectively, showing similar level of goodness of fit. In particular, among the top 20 genes selected using the Elem-OLS method with entrywise transformed Gram matrix in Yang, Lozano, and Ravikumar (2014), we found that five genes (AT1G57770, AT1G78670, AT3G56960, AT2G27820, and AT4G13700) selected by RANK are included in such a list of top 20 genes, and three genes (AT1G57770, AT1G78670, and AT2G27820) picked by RANK$_+$ are contained in the same list.

To gain some scientific insights into the selected genes, we conducted gene ontology (GO) enrichment analysis to interpret, from the biological point of view, the influence of selected genes on isoprenoid gene GGPPS11, which is known as a precursor to chloroplast, carotenoids, tocopherols, and abscisic acids. Specifically, in the enrichment test of GO biological process, gene AT1G57770 is involved in carotenoid biosynthetic process. In the GO cellular component enrichment test, genes AT4G38970 and AT5G19220 are located in chloroplast, chloroplast envelope, and chloroplast stroma; gene AT1G57770 is located in chloroplast and mitochondrion; and gene AT2G27820 is located in chloroplast, chloroplast stroma, and cytosol. The GO molecular function enrichment test shows that gene AT4G38970 has fructose-bisphosphate aldolase activity and gene AT1G57770 has carotenoid isomerase activity and oxidoreductase activity. These scientific insights in terms of biological process, cellular component, and molecular function suggest that the selected genes may have meaningful biological relationship with the target isoprenoid gene GGPPS11. See, for example, Horvath, Schaffer, and Wisman (2003), Ramel et al. (2009), and Wienkoop et al. (2004) for more discussions on these genes.

6. Discussions

Our analysis in this article reveals that the suggested RANK method exploiting the general framework of model-X knockoffs introduced in Candès et al. (2018) can asymptotically control the FDR in general high-dimensional nonlinear models with *unknown* covariate distribution. The robustness of the FDR control under *estimated* covariate distribution is enabled by imposing the Gaussian graphical structure on the covariates. Such a structural assumption has been widely employed to model the association networks among the covariates and extensively studied in the literature. Our method and theoretical results are powered by scalable large precision matrix estimation with statistical efficiency. It would be interesting to extend the robustness theory of the FDR control beyond Gaussian designs as well as for heavy-tailed data and dependent observations.

Our work also provides a first attempt to the power analysis for the model-X knockoffs framework. The nontrivial technical analysis establishes that RANK can have asymptotic power one in high-dimensional linear model setting when the Lasso is used for sparse regression. It would be interesting to extend the power analysis for RANK with a wide class of sparse regression and feature screening methods including SCAD, SIS, and many other concave regularization methods Fan and Li (2001), Fan and Lv (2008), Fan and Fan (2008), and Fan and Lv (2013). Though

more challenging, it is also important to investigate the power property for RANK beyond linear models. The power analysis in general high-dimensional nonlinear models is highly challenging for several reasons. First, the minimum signal strength needs to be characterized precisely in the power analysis. Yet unlike the beta-min measure in the linear model, there lacks a popularly accepted measure with explicit formula on the minimum signal strength in general high-dimensional nonlinear models. Second, the estimation error associated with each covariate plays an important role in the power analysis. However, in general nonlinear models it is unclear how to disentangle the individual estimation error corresponding to each covariate. Third, the knockoffs procedure builds on some underlying variable selection method, which itself is highly challenging both empirically and theoretically in general high-dimensional nonlinear models.

Our RANK procedure uses the idea of data splitting, which plays an important role in our technical analysis. Our numerical examples, however, suggest that data splitting is more of a technical assumption than a practical necessity. It would be interesting to develop theoretical guarantees for RANK without data splitting. These extensions are interesting topics for future research.

Supplementary Materials

The supplementary material contains additional technical details for the proofs of Lemmas 3–8.

Acknowledgment

Funding

Appendix A: Appendix: Proofs of Main Results

We provide the proofs of Theorems 1–3, Propositions 1 and 2, and Lemmas 1 and 2 in this appendix. Additional technical details for the proofs of Lemmas 3–8 are included in the supplementary material. To ease the technical presentation, we first introduce some notation. Let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ be the smallest and largest eigenvalues of a symmetric matrix. For any matrix $\mathbf{A} = (a_{ij})$, denote by $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$, $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, $\|\mathbf{A}\|_2 = \Lambda_{\max}^{1/2}(\mathbf{A}^T \mathbf{A})$, and $\|\mathbf{A}\|_F = [\mathrm{tr}(\mathbf{A}^T \mathbf{A})]^{1/2}$ the matrix ℓ_1-norm, entrywise maximum norm, spectral norm, and Frobenius norm, respectively. For any set $\mathcal{S} \subset \{1, \ldots, p\}$, we use $\mathbf{A}_{\mathcal{S}}$ to represent the submatrix of $\mathbf{A}$ formed by columns in set $\mathcal{S}$ and $\mathbf{A}_{\mathcal{S},\mathcal{S}}$ to denote the principal submatrix formed by columns and rows in set $\mathcal{S}$.

A.1. Proofs of Lemma 1 and Theorem 1

Observe that the choice of $\widetilde{\mathcal{S}} = \{1, \ldots, p\}$ certainly satisfies the sure screening property. We see that Lemma 1 and Theorem 1 are

specific cases of Lemma 6 in Section B.4 of supplementary materials and Theorem 3, respectively. Thus, we only prove the latter ones.

A.2. Proof of Proposition 1

In this proof, we will consider $\boldsymbol{\Omega}$ and $\mathcal{S}$ as deterministic parameters and focus only on the second half of sample $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ used in FDR control. Thus, we will drop the superscripts in $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ whenever there is no confusion. For a given precision matrix $\boldsymbol{\Omega}$, the matrix of knockoff variables

$$\widetilde{\mathbf{X}}^{\boldsymbol{\Omega}} = [\widetilde{\mathbf{x}}_1^{\boldsymbol{\Omega}}, \ldots, \widetilde{\mathbf{x}}_n^{\boldsymbol{\Omega}}]^T$$

can be generated using Equation (12) with $\boldsymbol{\Omega}_0$ replaced by $\boldsymbol{\Omega}$. Here, we use the superscript $\boldsymbol{\Omega}$ to emphasize the dependence of knockoffs matrix on $\boldsymbol{\Omega}$. Recall that for a given set $\mathcal{S}$ with $k = |\mathcal{S}|$, we calculate the knockoff statistics W_j's using $(\mathbf{y}, \mathbf{X}_\mathcal{S}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}})$. Thus, the FDR function can be written as

$$\text{FDR}_n(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[\text{FDP}_n(\mathbf{y}, \mathbf{X}_\mathcal{S}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}})] = \mathbb{E}[g_{1,n}(\mathbf{X}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}})], \quad (\text{A.1})$$

where $g_{1,n}(\mathbf{X}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}}) = \mathbb{E}[\text{FDP}_n(\mathbf{y}, \mathbf{X}_\mathcal{S}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}}) | \mathbf{X}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}}]$. It is seen that the function $g_{1,n}$ is the conditional FDP when knockoff variables $\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}}$, $1 \leq i \leq n$, are simulated using $\boldsymbol{\Omega}$ and only variables in set $\mathcal{S}$ are used to construct knockoff statistics W_j. We want to emphasize that since given $\mathbf{X}$ the response $\mathbf{y}$ is independent of $\widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}}$, the functional form of $g_{1,n}$ is free of the matrix $\boldsymbol{\Omega}$ used to generate knockoff variables.

Using the technical arguments in Candès et al. (2018), we can show that $\text{FDR}_n(\boldsymbol{\Omega}_0, \mathcal{S}) \leq q$ for any sample size n and all subsets $\mathcal{S} \subset \{1, \ldots, p\}$ that are independent of the original data $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ used in the knockoffs procedure. Observe that the only difference between $\text{FDR}_n(\boldsymbol{\Omega}, \mathcal{S})$ and $\text{FDR}_n(\boldsymbol{\Omega}_0, \mathcal{S})$ is that different precision matrices are used to generate knockoff variables. We restrict ourselves to the following data generating scheme

$$\widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}} = (\mathbf{C}^{\boldsymbol{\Omega}})^T \mathbf{x}_i + \mathbf{B}^{\boldsymbol{\Omega}} \mathbf{z}_i, \quad i = 1, \ldots, n,$$

where $\mathbf{C}^{\boldsymbol{\Omega}} = \mathbf{I}_p - \boldsymbol{\Omega}\text{diag}\{\mathbf{s}\}$, $\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}_p) \in \mathbb{R}^p$ are iid normal random vectors that are independent of $\mathbf{x}_i$'s, and $\mathbf{B}^{\boldsymbol{\Omega}} = (2\text{diag}\{\mathbf{s}\} - \text{diag}\{\mathbf{s}\}\boldsymbol{\Omega}\text{diag}\{\mathbf{s}\})^{1/2}$. For simplicity, write $\widetilde{\mathbf{x}}_i^{(0)} = \widetilde{\mathbf{x}}_i^{\boldsymbol{\Omega}_0}$, $\mathbf{B}_0 = \mathbf{B}^{\boldsymbol{\Omega}_0}$, and $\mathbf{C}_0 = \mathbf{C}^{\boldsymbol{\Omega}_0}$, that is, the matrices corresponding to the oracle case. Then restricted to set $\mathcal{S}$,

$$\widetilde{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}} = (\mathbf{C}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{x}_i + (\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{z}_i, \quad \widetilde{\mathbf{x}}_{i,\mathcal{S}}^{(0)} = \mathbf{C}_{0,\mathcal{S}}^T \mathbf{x}_i + (\mathbf{B}_{0,\mathcal{S}})^T \mathbf{z}_i,$$

where the subscript $\mathcal{S}$ means the submatrix (subvector) formed by columns (components) in set $\mathcal{S}$. We want to make connections between $\widetilde{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}}$ and $\widetilde{\mathbf{x}}_{i,\mathcal{S}}^{(0)}$. To this end, construct

$$\check{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}} = (\mathbf{C}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{x}_i + \widetilde{\mathbf{B}}^T \mathbf{B}_{0,\mathcal{S}}^T \mathbf{z}_i, \quad (\text{A.2})$$

where $\widetilde{\mathbf{B}} = (\mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})^{-1/2}((\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^{1/2}$. Then it is seen that $(\mathbf{x}_i, \widetilde{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}})$ and $(\mathbf{x}_i, \check{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}})$ have identical joint distribution. Although $\check{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}}$ cannot be calculated in practice for a given $\boldsymbol{\Omega}$ due to its dependency on $\boldsymbol{\Omega}_0$, the random vector $\check{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}}$ acts as a proxy of $\widetilde{\mathbf{x}}_{i,\mathcal{S}}^{\boldsymbol{\Omega}}$ in studying the FDR function. In fact, by construction (A.1) can be further written as

$$\text{FDR}_n(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[g_{1,n}(\mathbf{X}, \widetilde{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}})] = \mathbb{E}[g_{1,n}(\mathbf{X}, \check{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}})], \quad (\text{A.3})$$

where $\check{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}} = [\check{\mathbf{x}}_{1,\mathcal{S}}^{\boldsymbol{\Omega}}, \ldots, \check{\mathbf{x}}_{n,\mathcal{S}}^{\boldsymbol{\Omega}}]^T$.

Observe that the randomness in both $\widetilde{\mathbf{X}}_\mathcal{S}^{(0)}$ and $\check{\mathbf{X}}_\mathcal{S}^{\boldsymbol{\Omega}}$ is fully determined by the same random matrices $\mathbf{X}$ and $\mathbf{Z}\mathbf{B}_{0,\mathcal{S}}$, which are independent of each other and whose rows are iid copies from $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ and

$N(\mathbf{0}, \mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})$, respectively. For this reason, we can rewrite the FDR function in Equation (A.3) as

$$\text{FDR}_n(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[g_{1,n}(\mathbf{X}_{\text{aug}}^{\mathcal{S}} \mathbf{H}^{\boldsymbol{\Omega}})],$$

where $\mathbf{X}_{\text{aug}}^{\mathcal{S}} = [\mathbf{X}, \widetilde{\mathbf{X}}_{0,\mathcal{S}}] = [\mathbf{X}, \mathbf{X}\mathbf{C}_{0,\mathcal{S}} + \mathbf{Z}\mathbf{B}_{0,\mathcal{S}}] \in \mathbb{R}^{n \times (p+k)}$ is the augmented matrix collecting columns of $\mathbf{X}$ and $\widetilde{\mathbf{X}}_{0,\mathcal{S}}$, and

$$\mathbf{H}^{\boldsymbol{\Omega}} = \begin{pmatrix} \mathbf{I}_p & \mathbf{C}_\mathcal{S}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}\widetilde{\mathbf{B}} \\ \mathbf{0} & \widetilde{\mathbf{B}} \end{pmatrix},$$

which completes the proof of Proposition 1.

A.3. Lemma 2 and Its Proof

Lemma 2. Assume that $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(a_n)$ with $a_n \to 0$ some deterministic sequence and all the notation the same as in Proposition 1. If $\Lambda_{\min}\{2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})\} \geq c_0$ and $\Lambda_{\max}(\boldsymbol{\Sigma}_0) \leq c_0^{-1}$ for some constant $c_0 > 0$, then it holds that

$$\|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_2 \leq c_1 \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(a_n),$$

where $\widetilde{\mathbf{B}}$ is given in Equation (A.2) and $c_1 > 0$ is some uniform constant independent of set $\mathcal{S}$.

Proof. We use C to denote some generic positive constant whose value may change from line to line. First note that

$$(\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}} - \mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}} = -(\text{diag}(\mathbf{s})(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0)\text{diag}(\mathbf{s}))_{\mathcal{S},\mathcal{S}}. \quad (\text{A.4})$$

Further, since $\boldsymbol{\Sigma}_0 - 2^{-1}\text{diag}(\mathbf{s})$ is positive definite it follows that $\|\mathbf{s}\|_\infty \leq 2\Lambda_{\max}(\boldsymbol{\Sigma}_0) \leq 2c_0^{-1}$. Thus, it holds that

$$\|(\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}} - \mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}}\|_2 \leq C\|(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0)_{\mathcal{S},\mathcal{S}}\|_2$$
$$\leq C\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(a_n).$$

For n large enough, by the triangle inequality we have

$$\Lambda_{\min}((\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}}) \geq \Lambda_{\min}(\mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})$$
$$+ \Lambda_{\min}((\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}} - \mathbf{B}_{0,\mathcal{S}}^T \mathbf{B}_{0,\mathcal{S}})$$
$$\geq \Lambda_{\min}(\mathbf{B}_0^T \mathbf{B}_0) - O(a_n)$$
$$= \Lambda_{\min}(2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})) - O(a_n)$$
$$\geq c_0/2.$$

In addition, $\Lambda_{\min}((\mathbf{B}_{0,\mathcal{S}})^T \mathbf{B}_{0,\mathcal{S}}) \geq \Lambda_{\min}((\mathbf{B}_0)^T \mathbf{B}_0) = \Lambda_{\min}(2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})) \geq c_0/2$. The above two inequalities together with Lemma 2.2 in Schmitt (1992) entail that

$$\left\|((\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^{1/2} - ((\mathbf{B}_{0,\mathcal{S}})^T \mathbf{B}_{0,\mathcal{S}})^{1/2}\right\|_2$$
$$\leq (\sqrt{c_0/2} + \sqrt{c_0/2})^{-1}\|(\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}} - (\mathbf{B}_{0,\mathcal{S}})^T \mathbf{B}_{0,\mathcal{S}}\|_2$$
$$\leq C\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(a_n), \quad (\text{A.5})$$

where the last step is because of Equation (A.4). Thus, it follows that

$$\|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_2 \leq \left\|((\mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^T \mathbf{B}_\mathcal{S}^{\boldsymbol{\Omega}})^{1/2} - ((\mathbf{B}_{0,\mathcal{S}})^T \mathbf{B}_{0,\mathcal{S}})^{1/2}\right\|_2$$
$$\times \|((\mathbf{B}_{0,\mathcal{S}})^T \mathbf{B}_{0,\mathcal{S}})^{-1/2}\|_2$$
$$\leq C\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 \Lambda_{\min}^{-1/2}((\mathbf{B}_0)^T \mathbf{B}_0) \leq C\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2, \quad (\text{A.6})$$

where the last step comes from assumption $\Lambda_{\min}(\mathbf{B}_0^T \mathbf{B}_0) = \Lambda_{\min}(2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})) \geq c_0$. This concludes the proof of Lemma 2.

A.4. Proof of Theorem 2

We now proceed to prove Theorem 2 with the aid of Lemma 2 in Section A.3. We use the same notation as in the proof of Proposition 1 and use $C > 0$ to denote a generic constant whose value may change from line to line.

We start with proving (15). By Condition 4, we have

$$|\text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S}) - \text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}_0}, \mathcal{S})| \leq L\|\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0}\|_F, \qquad (A.7)$$

where the constant L is uniform over all $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\| \leq C_2 a_n$ and $|\mathcal{S}| \leq K_n$. Denote by $k = |\mathcal{S}|$. By the definition of $\mathbf{H}^{\boldsymbol{\Omega}}$, it holds that

$$\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0} = \mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{I}_{p+k} = \begin{pmatrix} \mathbf{0} & \mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}\widetilde{\mathbf{B}} \\ \mathbf{0} & \widetilde{\mathbf{B}} - \mathbf{I}_k \end{pmatrix}.$$

By the definition and matrix norm inequality, we deduce

$$\begin{aligned}
\|\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0}\|_F &= \|\mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}\widetilde{\mathbf{B}}\|_F + \|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_F \\
&\leq \sqrt{k}\|\mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}\widetilde{\mathbf{B}}\|_2 + \sqrt{k}\|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_2 \\
&\leq \sqrt{K_n}\Big(\|\mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} - \mathbf{C}_{0,\mathcal{S}}\|_2 + \|\mathbf{C}_{0,\mathcal{S}}(\widetilde{\mathbf{B}} - \mathbf{I}_k)\|_2 \\
&\qquad + \|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_2\Big) \\
&\leq \sqrt{K_n}\Big((1 + \|\mathbf{C}_{0,\mathcal{S}}\|_2)\|\widetilde{\mathbf{B}} - \mathbf{I}_k\|_2 + \|\mathbf{C}^{\boldsymbol{\Omega}} - \mathbf{C}_0\|_2\Big).
\end{aligned}$$

Since $\boldsymbol{\Sigma}_0 - 2^{-1}\text{diag}(\mathbf{s})$ is positive definite, it follows that $s_j \leq 2\Lambda_{\max}(\Sigma_0) \leq 2/\Lambda_{\min}(\boldsymbol{\Omega}_0) \leq C$. Thus, $\|\mathbf{C}_{0,\mathcal{S}}\|_2 \leq \|\mathbf{C}_0\|_2 = \|\mathbf{I} - \boldsymbol{\Omega}_0\text{diag}(\mathbf{s})\|_2 \leq 1 + \|\boldsymbol{\Omega}_0\|_2\|\text{diag}(\mathbf{s})\|_2 \leq C$. This along with $\|\mathbf{C}^{\boldsymbol{\Omega}} - \mathbf{C}_0\|_2 = \|(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0)\text{diag}(\mathbf{s})\|_2 \leq Ca_n$ and Lemma 2 entails that $\|\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0}\|_F$ can be further bounded as

$$\|\mathbf{H}^{\boldsymbol{\Omega}} - \mathbf{H}^{\boldsymbol{\Omega}_0}\|_F \leq C\sqrt{K_n}a_n.$$

Combining the above result with Equation (A.7) leads to

$$\sup_{\{|\mathcal{S}|\leq K_n, \|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|\leq Ca_n\}} |\text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S}) - \text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}_0}, \mathcal{S})| \leq O(\sqrt{K_n}a_n), \tag{A.8}$$

which completes the proof of Equation (15).

We next establish the FDR control for RANK. By Condition 7, the event $\mathcal{E}_0 = \{\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_2 \leq C_2 a_n\}$ occurs with probability at least $1 - O(p^{-c_1})$. Since $\widehat{\boldsymbol{\Omega}}$ and $\widetilde{\mathcal{S}}$ are estimated from independent subsample $(\mathbf{X}^{(1)}, \mathbf{y}^{(1)})$, it follows from Equation (15) that

$$\begin{aligned}
&\left|\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}})|\mathcal{E}_0\big] - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widetilde{\mathbf{X}}_{0,\widetilde{\mathcal{S}}})|\mathcal{E}_0\big]\right| \\
&\leq \sup_{|\mathcal{S}|\leq K_n, \|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|\leq C_2 a_n} \left|\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\mathcal{S}}^{(2)}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}})|\mathcal{E}_0\big] \right. \\
&\qquad\qquad\qquad\qquad\qquad \left. - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\mathcal{S}}^{(2)}, \widetilde{\mathbf{X}}_{0,\mathcal{S}})|\mathcal{E}_0\big]\right| \tag{A.9} \\
&= \sup_{|\mathcal{S}|\leq K_n, \|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|\leq C_2 a_n} \left|\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\mathcal{S}}^{(2)}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}})\big] \right. \\
&\qquad\qquad\qquad\qquad\qquad \left. - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\mathcal{S}}^{(2)}, \widetilde{\mathbf{X}}_{0,\mathcal{S}})\big]\right| \tag{A.10} \\
&= \sup_{|\mathcal{S}|\leq K_n, \|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|\leq C_2 a_n} \left|\text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}}, \mathcal{S}) - \text{FDR}(\mathbf{H}^{\boldsymbol{\Omega}_0}, \mathcal{S})\right| \\
&\leq O(\sqrt{K_n}a_n). \tag{A.11}
\end{aligned}$$

Now note that by the property of conditional expectation, we have

$$\begin{aligned}
&\text{FDR}_n(\widehat{\boldsymbol{\Omega}}, \widetilde{\mathcal{S}}) - \text{FDR}_n(\boldsymbol{\Omega}_0, \widetilde{\mathcal{S}}) \\
&= \Big(\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}})|\mathcal{E}_0\big] - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widetilde{\mathbf{X}}_{0,\widetilde{\mathcal{S}}})|\mathcal{E}_0\big]\Big)\,\mathbb{P}(\mathcal{E}_0) \\
&\quad + \Big(\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}})|\mathcal{E}_0^c\big] - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widetilde{\mathbf{X}}_{0,\widetilde{\mathcal{S}}})|\mathcal{E}_0^c\big]\Big)\,\mathbb{P}(\mathcal{E}_0^c) \\
&\equiv I_1 + I_2.
\end{aligned}$$

Let us first consider term I_1. By Equation (A.11), it holds that

$$\begin{aligned}
|I_1| &\leq \left|\mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widehat{\mathbf{X}}_{\widetilde{\mathcal{S}}})|\mathcal{E}_0\big] - \mathbb{E}\big[\text{FDP}_n(\mathbf{X}_{\widetilde{\mathcal{S}}}^{(2)}, \widetilde{\mathbf{X}}_{0,\widetilde{\mathcal{S}}})|\mathcal{E}_0\big]\right| \\
&\leq O(\sqrt{K_n}a_n).
\end{aligned}$$

We next consider term I_2. Since FDP is always bounded between 0 and 1, we have

$$|I_2| \leq 2\mathbb{P}(\mathcal{E}_0^c) \leq O(p^{-c_1}).$$

Combining the above two results yields

$$\left|\text{FDR}_n(\widehat{\boldsymbol{\Omega}}, \widetilde{\mathcal{S}}) - \text{FDR}_n(\boldsymbol{\Omega}_0, \widetilde{\mathcal{S}})\right| \leq O(\sqrt{K_n}a_n) + O(p^{-c_1}).$$

This together with the result of $\text{FDR}_n(\boldsymbol{\Omega}_0, \widetilde{\mathcal{S}}) \leq q$ mentioned in the proof of Proposition 1 in Section A.2 completes the proof of Theorem 2.

A.5. Proof of Theorem 3

In this proof, we will drop the superscripts in $(\mathbf{X}^{(2)}, \mathbf{y}^{(2)})$ whenever there is no confusion. By the definition of power, for any given precision matrix $\boldsymbol{\Omega}$ and reduced model $\mathcal{S}$ the power can be written as

$$\text{Power}(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[f(\mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})],$$

where f is some function describing how the empirical power depends on the data. Note that $f(\mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})$ is a stochastic process indexed by $\boldsymbol{\Omega}$, and we care about the mean of this process. Our main idea is to construct another stochastic process indexed by $\boldsymbol{\Omega}$ which has the same mean but possibly different distribution. Then by studying the mean of this new stochastic process, we can prove the desired result.

We next provide more technical details of the proof. The proxy process is defined as

$$\check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}} = \mathbf{X}\mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}} + \mathbf{Z}\mathbf{B}_{0,\mathcal{S}}(\mathbf{B}_{0,\mathcal{S}}^T\mathbf{B}_{0,\mathcal{S}})^{-1/2}\Big((\mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}})^T\mathbf{B}_{\mathcal{S}}\Big)^{1/2}, \tag{A.12}$$

where $\mathbf{C}_{\mathcal{S}}^{\boldsymbol{\Omega}}$ is the submatrix of $\mathbf{C}^{\boldsymbol{\Omega}} = \mathbf{I}_p - \boldsymbol{\Omega}\text{diag}\{\mathbf{s}\}$, $\mathbf{B}_{\mathcal{S}}^{\boldsymbol{\Omega}}$ is the submatrix of $\mathbf{B}^{\boldsymbol{\Omega}} = \big(\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}\text{diag}(\mathbf{s})\big)^{1/2}$, and $\mathbf{B}_0 = \mathbf{B}^{\boldsymbol{\Omega}_0}$. It is easy to see that $\check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}$ and $\widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}$ defined using Equation (12) have the same distribution. Since $\mathbf{Z}$ is independent of $(\mathbf{X}, \mathbf{y})$, we can further conclude that $(\mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})$ and $(\mathbf{X}_{\mathcal{S}}, \check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})$ have the same joint distribution for each given $\boldsymbol{\Omega}$ and $\mathcal{S}$. Thus, the power function can be further written as

$$\text{Power}(\boldsymbol{\Omega}, \mathcal{S}) = \mathbb{E}[f(\mathbf{X}_{\mathcal{S}}, \widetilde{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})] = \mathbb{E}[f(\mathbf{X}_{\mathcal{S}}, \check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})].$$

Therefore, we only need to study the power of the knockoffs procedure based on the pseudo data $(\mathbf{X}_{\mathcal{S}}, \check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})$.

To simplify the technical presentation, we will slightly abuse the notation and still use $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S})$ to represent the Lasso solution based on pseudo data $(\mathbf{X}_{\mathcal{S}}, \check{\mathbf{X}}_{\mathcal{S}}^{\boldsymbol{\Omega}}, \mathbf{y})$. We will use c and C to denote some generic positive constants whose values may change from line to line. Define

$$\widetilde{\mathbf{G}} = \frac{1}{n}\widetilde{\mathbf{X}}_{\text{KO}}^T\widetilde{\mathbf{X}}_{\text{KO}} \in \mathbb{R}^{(2p)\times(2p)} \quad \text{and} \quad \widetilde{\boldsymbol{\rho}} = \frac{1}{n}\widetilde{\mathbf{X}}_{\text{KO}}^T\mathbf{y} \in \mathbb{R}^{2p} \tag{A.13}$$

with $\widetilde{\mathbf{X}}_{\text{KO}} = [\mathbf{X}, \check{\mathbf{X}}^{\boldsymbol{\Omega}}] \in \mathbb{R}^{n\times(2p)}$ the augmented design matrix. For any given set $\mathcal{S} \subset \{1, \ldots, p\}$ with $k = |\mathcal{S}|$, $(2p) \times (2p)$ matrix $\mathbf{A}$, and $(2p)$-vector $\mathbf{a}$, we will abuse the notation and denote by $\mathbf{A}_{\mathcal{S},\mathcal{S}} \in \mathbb{R}^{(2k)\times(2k)}$ the principal submatrix formed by columns and rows in set $\{j : j \in \mathcal{S} \text{ or } j - p \in \mathcal{S}\}$ and $\mathbf{a}_{\mathcal{S}} \in \mathbb{R}^{2k}$ the subvector formed by components in set $\{j : j \in \mathcal{S} \text{ or } j - p \in \mathcal{S}\}$. For any $p \times p$ matrix $\mathbf{B}$ (or p-vector $\mathbf{b}$), we define $\mathbf{B}_{\mathcal{S}}$ (or $\mathbf{b}_{\mathcal{S}}$) in the same way meaning that

columns (or components) in set $\mathcal{S}$ will be taken to form the submatrix (or subvector).

With the above notation, note that the Lasso solution $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{2p})^T = \widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S})$ restricted to variables in $\mathcal{S}$ can be obtained by setting $\widehat{\beta}_j = 0$ for $j \in \{1 \le j \le 2p : j \notin \mathcal{S} \text{ and } j - p \notin \mathcal{S}\}$ and minimizing the following objective function

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S}} = \arg\min_{\mathbf{b} \in \mathbb{R}^{2k}} \left\{ \frac{1}{2} \mathbf{b}^T \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}} \mathbf{b} - \widetilde{\boldsymbol{\rho}}_{\mathcal{S}}^T \mathbf{b} + \lambda \|\mathbf{b}\|_1 \right\}. \quad (\text{A.14})$$

By Proposition 2 in Section A.6, it holds that with probability at least $1 - c_3 p^{-c_3}$,

$$\sup_{\boldsymbol{\Omega} \in \mathcal{A}, |\mathcal{S}| \le K_n} \|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_{\mathbb{T}}\|_1 \le C_{\ell_1} s \lambda, \quad (\text{A.15})$$

where $\lambda = C_\lambda \{(\log p)/n\}^{1/2}$ with $C_\lambda > 0$ some constant and C_{ℓ_1} is some positive constant. By Condition 2 and the assumption $\lambda = C_\lambda \{(\log p)/n\}^{1/2}$, we have

$$\min_{j \in \mathcal{S}_0} |\beta_{0,j}| \ge \frac{\kappa_n \lambda}{C_\lambda}. \quad (\text{A.16})$$

Denote by $W_j^{\boldsymbol{\Omega}, \mathcal{S}}$ the LCD based on the above $\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S})$. Recall that by assumption, there are no ties in the magnitude of nonzero $W_j^{\boldsymbol{\Omega}, \mathcal{S}}$'s and no ties in the nonzero components of the Lasso solution with asymptotic probability one. Let $|W_{(1)}^{\boldsymbol{\Omega}, \mathcal{S}}| \ge \cdots \ge |W_{(p)}^{\boldsymbol{\Omega}, \mathcal{S}}|$ be the ordered knockoff statistics according to magnitude. Denote by j^* the index such that $|W_{(j^*)}^{\boldsymbol{\Omega}, \mathcal{S}}| = T$. Then by the definition of T, it holds that $-T < W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}} \le 0$. We next analyze the two cases of $W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}} = 0$ and $-T < W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}} < 0$ separately.

Case 1. For the case of $W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}} = 0$, $\widehat{\mathcal{S}}^{\boldsymbol{\Omega}} = \{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} > 0\}$, and $\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < -T\} = \{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\}$. Let $\widetilde{q}_n = \varphi C_{\ell_1} C_\lambda \kappa_n^{-1}$ with φ the positive solution to equation $\varphi^2 - \varphi - 1 = 0$ which is known as the golden ratio. If $|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\}| > \widetilde{q}_n s$, then we can prove that $T \le \kappa_n \lambda/(C_\lambda \varphi)$ using the same arguments as in Equation (A.21) since $\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < -T\} = \{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\}$. Thus, it reduces to Case 2 below and the arguments therein follow.

On the contrary, if $|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\}| \le \widetilde{q}_n s$ then we have

$$\left| \widehat{\mathcal{S}}^{\boldsymbol{\Omega}} \cap \mathcal{S}_0 \right| = \left| \text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \cap \mathcal{S}_0 \right| - \left| \{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\} \cap \mathcal{S}_0 \right|$$
$$\ge \left| \text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \cap \mathcal{S}_0 \right| - \widetilde{q}_n s \quad (\text{A.17})$$

since $\widehat{\mathcal{S}}^{\boldsymbol{\Omega}} = \text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \setminus \{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} < 0\}$. Let us now focus on $\left| \text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \cap \mathcal{S}_0 \right|$. We observe that

$$\text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \supset \{1, \ldots, p\} \setminus \mathcal{S}_1^{\boldsymbol{\Omega}}, \quad (\text{A.18})$$

where $\mathcal{S}_1^{\boldsymbol{\Omega}} = \{1 \le j \le p : \widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S}) = 0\}$. Meanwhile, note that in view of Equation (A.15) we have with probability at least $1 - c_3 p^{-c_3}$,

$$C_{\ell_1} s \lambda \ge \sup_{\boldsymbol{\Omega} \in \mathcal{A}, |\mathcal{S}| \le K_n} \|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_0\|_1$$
$$\ge \sup_{\boldsymbol{\Omega} \in \mathcal{A}, |\mathcal{S}| \le K_n} \sum_{j \in \mathcal{S}_1^{\boldsymbol{\Omega}} \cap \mathcal{S}_0} |\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \beta_{0,j}|$$
$$= \sum_{j \in \mathcal{S}_1^{\boldsymbol{\Omega}} \cap \mathcal{S}_0} |\beta_{0,j}| \ge |\mathcal{S}_1^{\boldsymbol{\Omega}} \cap \mathcal{S}_0| \min_{j \in \mathcal{S}_0} |\beta_{0,j}|.$$

By Equation (A.16), we can further deduce from the above inequality that

$$|\mathcal{S}_1^{\boldsymbol{\Omega}} \cap \mathcal{S}_0| \le C_{\ell_1} C_\lambda \kappa_n^{-1} s,$$

which together with $|\mathcal{S}_0| = s$ entails that

$$|(\{1, \ldots, p\} \setminus \mathcal{S}_1^{\boldsymbol{\Omega}}) \cap \mathcal{S}_0| \ge (1 - C_{\ell_1} C_\lambda \kappa_n^{-1}) s.$$

Combining this result with Equation (A.18) yields

$$\left| \text{supp}(\mathbf{W}^{\boldsymbol{\Omega}, \mathcal{S}}) \cap \mathcal{S}_0 \right| \ge \left| (\{1, \ldots, p\} \setminus \mathcal{S}_1^{\boldsymbol{\Omega}}) \cap \mathcal{S}_0 \right| \ge (1 - C_{\ell_1} C_\lambda \kappa_n^{-1}) s. \quad (\text{A.19})$$

Thus, in view of inequalities Equations (A.17) and (A.19), with probability at least $1 - c_3 p^{-c_3}$ it holds uniformly over all $\boldsymbol{\Omega} \in \mathcal{A}$ and $|\mathcal{S}| \le K_n$ that

$$\frac{|\widehat{\mathcal{S}}^{\boldsymbol{\Omega}} \cap \mathcal{S}_0|}{s} \ge 1 - C_{\ell_1} C_\lambda (\varphi + 1) \kappa_n^{-1}.$$

Case 2. We next consider the case of $-T < W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}} < 0$. By the definitions of T and j^*, we have

$$\frac{|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} \le -T\}| + 2}{|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} \ge T\}|} > q \quad (\text{A.20})$$

since otherwise we would reduce T to $|W_{(j^*+1)}^{\boldsymbol{\Omega}, \mathcal{S}}|$ to get the new smaller threshold with the criterion still satisfied. We next bound T using the results in Lemma 6 in Section B.4 of supplementary materials. Observe that Equation (A.20) and Lemma 6 lead to $|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} \le -T\}| > q|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} \ge T\}| - 2 \ge qcs - 2$ with asymptotic probability one. Moreover, when $W_j^{\boldsymbol{\Omega}, \mathcal{S}} \le -T$ we have $|\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S})| - |\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})| \le -T$ and thus $|\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})| \ge T$. Using Equation (A.15), we obtain

$$C_{\ell_1} s \lambda \ge \|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_{\mathbb{T}}\|_1 \ge \sum_{j: W_j^{\boldsymbol{\Omega}, \mathcal{S}} \le -T} |\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})|$$
$$\ge T|\{j : W_j^{\boldsymbol{\Omega}, \mathcal{S}} \le -T\}|. \quad (\text{A.21})$$

Combining these results leads to $C_{\ell_1} s \lambda \ge T(qcs - 2)$ and thus it holds that

$$T \le \frac{C_{\ell_1} s \lambda}{(qcs - 2)} \le \frac{\kappa_n \lambda}{C_\lambda \varphi} \quad (\text{A.22})$$

for large enough n since $\kappa_n \to \infty$ as $n \to \infty$ and C_λ is some positive constant.

We now proceed to prove the theorem by showing that Type II error is small. In light of Equation (A.15), we derive

$$C_{\ell_1} s \lambda \ge \|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_{\mathbb{T}}\|_1 = \sum_{j=1}^{p} \left[|\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \beta_{0,j}| \right.$$
$$\left. + |\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})| \right]$$
$$\ge \sum_{j \in \mathcal{S}_0 \cap (\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c} \left[|\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \beta_{0,j}| + |\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})| \right]$$
$$\ge \sum_{j \in \mathcal{S}_0 \cap (\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c} \left[|\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \beta_{0,j}| + |\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S})| - T \right]$$

since $|\widehat{\beta}_{j+p}(\lambda; \boldsymbol{\Omega}, \mathcal{S})| \ge |\widehat{\beta}_j(\lambda; \boldsymbol{\Omega}, \mathcal{S})| - T$ when $j \in (\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c$. Using the triangle inequality and noting that $|\beta_{0,j}| \ge C_\lambda^{-1} \lambda \kappa_n$ for $j \in \mathcal{S}_0$, we can conclude that

$$C_{\ell_1} s \lambda \ge \sum_{j \in \mathcal{S}_0 \cap (\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c} (|\beta_{0,j}| - T) \ge (C_\lambda^{-1} \lambda \kappa_n - T) \left| (\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c \cap \mathcal{S}_0 \right|.$$

Thus, it follows that

$$\frac{|\widehat{\mathcal{S}}^{\boldsymbol{\Omega}} \cap \mathcal{S}_0|}{s} = 1 - \frac{|(\widehat{\mathcal{S}}^{\boldsymbol{\Omega}})^c \cap \mathcal{S}_0|}{s} \geq 1 - \frac{C_{\ell_1}\lambda}{C_\lambda^{-1}\lambda\kappa_n - T}$$

$$= 1 - C_{\ell_1}C_\lambda\frac{\varphi}{\varphi - 1}\kappa_n^{-1}$$

uniformly over all $\boldsymbol{\Omega} \in \mathcal{A}$ and $|\mathcal{S}| \leq K_n$ since $T \leq \kappa_n\lambda/(C_\lambda\varphi)$.

Combining the above two scenarios, we have shown that with asymptotic probability one, uniformly over all $\boldsymbol{\Omega} \in \mathcal{A}$ and $|\mathcal{S}| \leq K_n$ it holds that with probability at least $1 - c_3 p^{-c_3}$,

$$\frac{|\widehat{\mathcal{S}}^{\boldsymbol{\Omega}} \cap \mathcal{S}_0|}{s} \geq 1 - C_{\ell_1}C_\lambda(\varphi + 1)\kappa_n^{-1} \qquad (A.23)$$

since $\varphi + 1 = \varphi/(\varphi - 1)$ by the definition of φ. This along with the assumption $\mathbb{P}\{\widehat{\boldsymbol{\Omega}} \in \mathcal{A}\} \geq 1 - c_2 p^{-c_2}$ in Condition 7 gives

$$\text{Power}(\widehat{\boldsymbol{\Omega}}, \widetilde{\mathcal{S}}) = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}}^{\widehat{\boldsymbol{\Omega}}} \cap \mathcal{S}_0|}{s}\right] \geq \mathbb{E}\left[\frac{|\widehat{\mathcal{S}}^{\widehat{\boldsymbol{\Omega}}} \cap \mathcal{S}_0|}{s}\bigg|\widehat{\boldsymbol{\Omega}} \in \mathcal{A}\right]\mathbb{P}\{\widehat{\boldsymbol{\Omega}} \in \mathcal{A}\}$$

$$\geq \left[1 - C_{\ell_1}C_\lambda(\varphi + 1)\kappa_n^{-1}\right](1 - c_3 p^{-c_3})(1 - c_2 p^{-c_2})$$

$$= 1 - C_{\ell_1}C_\lambda(\varphi + 1)\kappa_n^{-1} - c_2 p^{-c_2} - c_3 p^{-c_3} + o(\kappa_n^{-1})$$

$$= 1 - o(1),$$

which concludes the proof of Theorem 3.

A.6. Proposition 2 and Its Proof

Proposition 2. Assume that Conditions 1 and 6 hold, the smallest eigenvalue of $2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})$ is positive and bounded away from 0, and $\lambda = C_\lambda\{(\log p)/n\}^{1/2}$ with $C_\lambda > 0$ some constant. Let $\boldsymbol{\beta}_{\mathbb{T}} = (\boldsymbol{\beta}_0^T, 0, \ldots, 0)^T \in \mathbb{R}^{2p}$ be the expanded vector of true regression coefficient vector. If $[(L_p + L'_p)^{1/2} + K_n^{1/2}]a_n = o(1)$ and $s\{a_n + L'_p[(\log p)/n]^{1/2} + [K_n(\log p)/n]^{1/2}\} = o(1)$, then with probability at least $1 - c_3 p^{-c_3}$,

$$\sup_{\boldsymbol{\Omega} \in \mathcal{A}, |\mathcal{S}| \leq K_n}\|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_{\mathbb{T}}\|_1 = C_{\ell_1}s\lambda$$

$$\text{and} \quad \sup_{\boldsymbol{\Omega} \in \mathcal{A}, |\mathcal{S}| \leq K_n}\|\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S}) - \boldsymbol{\beta}_{\mathbb{T}}\|_2 = C_{\ell_2}s^{1/2}\lambda,$$

where $\widehat{\boldsymbol{\beta}}(\lambda; \boldsymbol{\Omega}, \mathcal{S})$ is defined in the proof of Theorem 3 in Section A.5 and c_3, c_4, C_{ℓ_1}, and C_{ℓ_2} are all positive constants.

Proof. We adopt the same notation as used in the proof of Theorem 3 in Section A.5. Let us introduce some key events which will be used in the technical analysis. Define

$$\mathcal{E}_3 = \left\{\sup_{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\| \leq C_2 a_n, |\mathcal{S}| \leq K_n}\|\widetilde{\boldsymbol{\rho}}_{\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_\infty \leq \lambda_0\right\}, \quad (A.24)$$

$$\mathcal{E}_4 = \left\{\sup_{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 \leq C_2 a_n, |\mathcal{S}| \leq K_n}\|\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}} - \mathbf{G}_{\mathcal{S},\mathcal{S}}\|_{\max} \leq C_5 a_{2,n}\right\}, \quad (A.25)$$

where $\lambda_0 = C_4\sqrt{(\log p)/n}$ and $a_{2,n} = a_n + (L'_p + K_n)\sqrt{(\log p)/n}$ with $C_4, C_5 > 0$ some constants. Then by Lemma 4 and 7 in Sections B.2 and B.5 of supplementary materials,

$$\mathbb{P}(\mathcal{E}_3 \cap \mathcal{E}_4) = 1 - c_3 p^{-c_3} \qquad (A.26)$$

for some constant $c_3 > 0$. Hereafter, we will condition on the event $\mathcal{E}_3 \cap \mathcal{E}_4$.

Since $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$ is the minimizer of the objective function in Equation (A.14), we have

$$\frac{1}{2}\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \widetilde{\boldsymbol{\rho}}_{\mathcal{S}}^T\widehat{\boldsymbol{\beta}}_{\mathcal{S}} + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$$

$$\leq \frac{1}{2}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}} - \widetilde{\boldsymbol{\rho}}_{\mathcal{S}}^T\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}} + \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1.$$

Some routine calculations lead to

$$\frac{1}{2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}})^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}) + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1$$

$$\leq -\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}} + \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}$$

$$+ \widetilde{\boldsymbol{\rho}}_{\mathcal{S}}^T(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}) + \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1$$

$$= (\widetilde{\boldsymbol{\rho}}_{\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}})^T(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}) + \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1 \qquad (A.27)$$

$$\leq \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1\|\widetilde{\boldsymbol{\rho}}_{\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_\infty + \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1. \qquad (A.28)$$

Let $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathbb{T}}$. Then we can simplify Equation (A.28) as

$$\frac{1}{2}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1 \leq \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1\|\widetilde{\boldsymbol{\rho}}_{\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_\infty$$

$$+ \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1$$

$$\leq \lambda_0\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + \lambda\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1. \qquad (A.29)$$

Observe that $\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}}\|_1 = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0}\|_1 + \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1$ and $\|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1 = \|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1 + \|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}\setminus\mathcal{S}_0}\|_1 = \|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1$ with $\mathcal{S}_0$ the support of true regression coefficient vector. Then it follows from $\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1 \geq \|\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1 - \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0}\|_1$ that

$$\frac{1}{2}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 \leq \lambda_0\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + \lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1.$$

Denote by $\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}_0} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}_0}\|_1$ and $\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}\setminus\mathcal{S}_0} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}\setminus\mathcal{S}_0}\|_1 = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1$. Then we can further deduce

$$\frac{1}{2}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} + \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 \leq \lambda_0\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1 + \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1$$

$$= \lambda_0\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 + \lambda_0\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 + \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1;$$

that is,

$$\frac{1}{2}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} + (\lambda - \lambda_0)\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 \leq (\lambda + \lambda_0)\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1. \qquad (A.30)$$

When $\lambda \geq 2\lambda_0$, it holds that

$$\frac{1}{2}\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} + \frac{\lambda}{2}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 \leq \frac{3\lambda}{2}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1. \qquad (A.31)$$

Since $\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} \geq 0$, we obtain the basic inequality

$$\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1 \leq 3\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 \qquad (A.32)$$

on event $\mathcal{E}_3$. It follows from Equation (A.31) that

$$\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T\mathbf{G}_{\mathcal{S},\mathcal{S}}\widehat{\boldsymbol{\delta}}_{\mathcal{S}} \leq 3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 + \widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T(\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}})\widehat{\boldsymbol{\delta}}_{\mathcal{S}} \qquad (A.33)$$

with

$$\mathbf{G} = \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_0 - \text{diag}(\mathbf{s}) \\ \boldsymbol{\Sigma}_0 - \text{diag}(\mathbf{s}) & \boldsymbol{\Sigma}_0 \end{pmatrix}.$$

With some matrix calculations, we can show that

$$\Lambda_{\min}(\mathbf{G}) \geq \Lambda_{\min}(\boldsymbol{\Sigma}_0)\Lambda_{\min}\{2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})\} \geq C,$$

since both $\boldsymbol{\Sigma}_0$ and $2\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\boldsymbol{\Omega}_0\text{diag}(\mathbf{s})$ have eigenvalues bounded away from 0. Thus, the left hand side of Equation (A.33) can be bounded from below by $c_0 C\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2$.

It remains to bound the right hand side of Equation (A.33). For the first term, it follows from the Cauchy–Schwarz inequality that

$$3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 \leq 3\lambda\sqrt{s}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_2 \leq 3\lambda\sqrt{s}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2. \tag{A.34}$$

For the last term $\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T(\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}})\widehat{\boldsymbol{\delta}}_{\mathcal{S}}$ on the right hand of Equation (A.33), by conditioning on event $\mathcal{E}_4$ and using the Cauchy–Schwarz inequality, the triangle inequality, and the basic inequality Equation (A.32) we can obtain

$$
\begin{aligned}
\left|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}^T(\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}})\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\right| &\leq \|\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\|_{\max}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_1^2 \\
&\leq \|\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\|_{\max}(\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1 + \|\widehat{\boldsymbol{\delta}}_{\mathcal{S}\setminus\mathcal{S}_0}\|_1)^2 \\
&\leq 16\|\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\|_{\max}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_1^2 \\
&\leq 16s\|\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\|_{\max}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}_0}\|_2^2 \\
&\leq 16s\|\mathbf{G}_{\mathcal{S},\mathcal{S}} - \widetilde{\mathbf{G}}_{\mathcal{S},\mathcal{S}}\|_{\max}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2 \\
&\leq 16C_5 sa_{2,n}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2.
\end{aligned}
$$

Combining the above results, we can reduce inequality Equation (A.33) to

$$C\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2 \leq 3\lambda\sqrt{s}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2 + 16C_5 sa_{2,n}\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2^2.$$

Since $sa_{2,n} \to 0$, there exists a positive constant C_{ℓ_2} such that it holds for n large enough that

$$\|\widehat{\boldsymbol{\delta}}_{\mathcal{S}}\|_2 = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_2 = C_{\ell_2}\sqrt{s}\lambda.$$

Further, by Equation (A.34) we have

$$\|\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathbb{T},\mathcal{S}}\|_1 = C_{\ell_1} s\lambda$$

for some constant C_{ℓ_1} and for large enough n. Note that by definition, $\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}$ and $\boldsymbol{\beta}_{\mathbb{T},\mathcal{S}^c} = \mathbf{0}$. Therefore, summarizing the above results completes the proof of Proposition 2. ∎

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006), "Adapting to Unknown Sparsity by Controlling the False Discovery Rate," *The Annals of Statistics*, 34, 584–653. [362]

Barber, R. F., and Candès, E. J. (2015), "Controlling the False Discovery Rate via Knockoffs," *The Annals of Statistics*, 43, 2055–2085. [363,364,367,369]

——— (2016), "A Knockoff Filter for High-Dimensional Selective Inference," arXiv:1602.03574. [363,364,367,369]

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, 57, 289–300. [362,363]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [362]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [365]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Berlin: Springer. [362,365]

Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'ModelX' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [363,364,365,366,367,373,374]

Chen, M., Ren, Z., Zhao, H., and Zhou, H. H. (2016), "Asymptotically Normal and Efficient Estimation of Covariate-Adjusted Gaussian Graphical Model," *Journal of the American Statistical Association*, 111, 394–406. [369]

Chouldechova, A., and Hastie, T. (2015), "Generalized Additive Model Selection," arXiv:1506.03850. [369,370]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [362]

Efron, B. (2007a), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [362]

——— (2007b), "Size, Power and False Discovery Rates," *The Annals of Statistics*, 35, 1351–1377. [362]

Efron, B., and Tibshirani, R. (2002), "Empirical Bayes Methods and False Discovery Rates for Microarrays," *Genetic Epidemiology*, 23, 70–86. [362]

Engle, R., Granger, C., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320. [369]

Fan, J., and Fan, Y. (2008), "High-Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [367,373]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall/CRC. [370]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [367]

Fan, J., Hall, P., and Yao, Q. (2007), "To How Many Simultaneous Hypothesis Tests Can Normal, Student's t or Bootstrap Calibration be Applied?," *Journal of the American Statistical Association*, 102, 1282–1288. [362]

Fan, J., Han, X., and Gu, W. (2012), "Estimating False Discovery Proportion Under Arbitrary Covariance Dependence," *Journal of American Statistical Association*, 107, 1019–1035. [362]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [373]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion)," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [363,367,373]

——— (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space (invited review article)," *Statistica Sinica*, 20, 101–148. [362]

Fan, J., Samworth, R. J., and Wu, Y. (2009), "Ultrahigh Dimensional Variable Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [367]

Fan, Y., Demirkaya, E., and Lv, J. (2017), "Nonuniformity of p-values Can Occur Early in Diverging Dimensions," arXiv:1705.03604. [363]

Fan, Y., and Fan, J. (2011), "Testing and Detecting Jumps Based on a Discretely Observed Process," *Journal of Econometrics*, 164, 331–344. [362]

Fan, Y., Kong, Y., Li, D., and Zheng, Z. (2015), "Innovated Interaction Screening for High-Dimensional Nonlinear Classification," *The Annals of Statistics*, 43, 1243–1272. [368]

Fan, Y., and Lv, J. (2013), "Asymptotic Equivalence of Regularization Methods in Thresholded Parameter Space," *Journal of the American Statistical Association*, 108, 1044–1061. [367,373]

——— (2016), "Innovated Scalable Efficient Estimation in Ultra-Large Gaussian Graphical Models," *The Annals of Statistics*, 44, 2098–2126. [369,370]

Hall, P., and Wang, Q. (2010), "Strong Approximations of Level Exceedances Related to Multiple Hypothesis Testing," *Bernoulli*, 16, 418–434. [362]

Härdle, W., Liang, H., and Gao, J. T. (2000), *Partially Linear Models*, Heidelberg: Springer Physica-Verlag. [369]

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American statistical Association*, 84, 986–995. [369]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*. London: Chapman & Hall/CRC. [369]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Berlin: Springer. [362]

Horowitz, J. L. (2009), *Semiparametric and Nonparametric Methods in Econometrics*. Berlin: Springer. [369]

Horvath, D. P., Schaffer, R., and Wisman, E. (2003), "Identification of Genes Induced in Emerging Tillers of Wild Oat (*Avena fatua*) Using Arabidopsis Microarrays," *Weed Science*, 51, 503–508. [372,373]

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [363]

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120. [369]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Oxford University Press. [363,364]

Li, Q., and J. S. Racine (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton, NJ: Princeton University Press. [369]

Lin, Q., Zhao, Z., and Liu, J. S. (2016), "Sparse Sliced Inverse Regression for High Dimensional Data," arXiv:1611.06655. [370]

Liu, W., and Shao, Q.-M. (2014), "Phase Transition and Regularized Bootstrap in Large-Scale *t*-tests With False Discovery Rate Control," *The Annals of Statistics*, 42, 2003–2025. [362]

Lv, J. (2013), "Impacts of High Dimensionality in Finite Samples," *The Annals of Statistics*, 41, 2236–2262. [367]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*. London: Chapman and Hall. [363]

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [369]

Meng, L., Sun, F., Zhang, X., and Waterman, M. S. (2011), "Sequence Alignment as Hypothesis Testing," *Journal of Computational Biology*, 18, 677–691. [362]

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006), "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data," *Bioinformatics*, 22, 1122–1129. [372]

Ramel, F., Sulmon, C., Bogard, M., Couée, I., and Gouesbet, G. (2009), "Differential Patterns of Reactive Oxygen Species and Antioxidative Mechanisms During Atrazine Injury and Sucrose-Induced Tolerance in Arabidopsis Thaliana Plantlets," *BMC Plant Biology*, 9, 1–18. [373]

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2009), "Spam: Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [369]

Ren, Z., Kang, Y., Fan, Y., and Lv, J. (2018), "Tuning-Free Heterogeneous Inference in Massive Networks," *Journal of the American Statistical Association*. [369]

Schäfer, J., and Strimmer, K. (2005), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Statistical Applications in Genetics and Molecular Biology*, 4, 1544–1615. [372]

Schmitt, B. A. (1992), "Perturbation Bounds for Matrix Square Roots and Pythagorean Sums," *Linear Algebra and Its Applications*, 174, 215–227. [374]

Shah, R. D., and Samworth, R. J. (2013), "Variable Selection With Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society*, Series B, 75, 55–80. [367]

Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481. [369]

Storey, J. D. (2002), A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, Series B, 64, 479–498. [362]

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Series B, 66, 187–205. [362]

Su, W., and Candès, E. J. (2016), "Slope Is Adaptive to Unknown Sparsity and Asymptotically Minimax," *The Annals of Statistics*, 44, 1038–1068. [362]

Sur, P., Chen, Y., and Candès, E. J. (2017), "The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square," arXiv:1706.01191. [363]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [364,365]

Wienkoop, S., Glinski, M., Tanaka, N., Tolstikov, V., Fiehn, O., and Weckwerth, W. (2004), "Linking Protein Fractionation With Multidimensional Monolithic Reversed-Phase Peptide Chromatography/Mass Spectrometry Enhances Protein Identification From Complex Mixtures Even in the Presence of Abundant Proteins," *Rapid Communications in Mass Spectrometry*, 18, 643–650. [373]

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelić, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Bühlmann, P. (2004), "Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in Arabidopsis Thaliana," *Genome Biology*, 5, R92. [372]

Wu, W. B. (2008), "On False Discovery Control Under Dependence," *The Annals of Statistics*, 36, 364–380. [362]

Yang, E., Lozano, A., and Ravikumar, P. (2014), "Elementary Estimators for High-Dimensional Linear Regression," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 388–396. [372,373]

Zhang, Y., and Liu, J. S. (2011), "Fast and Accurate Approximation to Significance Tests in Genome-Wide Association Studies," *Journal of the American Statistical Association*, 106, 846–857. [362]