DSSR

Discussion Paper No.92

IPAD: Stable Interpretable Forecasting with Knockoffs Inference

Yingying Fan, Jinchi Lv, Mahrad Sharifvaghefi
and Yoshimasa Uematsu

January 29, 2019

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

IPAD: Stable Interpretable Forecasting with Knockoffs Inference [*]

Yingying Fan[1], Jinchi Lv[1], Mahrad Sharifvaghefi[1] and Yoshimasa Uematsu[2]
University of Southern California[1] and Tohoku University[2]

January 29, 2019

Interpretability and stability are two important features that are desired in many contemporary big data applications arising in economics and finance. While the former is enjoyed to some extent by many existing forecasting approaches, the latter in the sense of controlling the fraction of wrongly discovered features which can enhance greatly the interpretability is still largely underdeveloped in the econometric settings. To this end, in this paper we exploit the general framework of model-X knockoffs introduced recently in Candès, Fan, Janson and Lv (2018), which is nonconventional for reproducible large-scale inference in that the framework is completely free of the use of p-values for significance testing, and suggest a new method of intertwined probabilistic factors decoupling (IPAD) for stable interpretable forecasting with knockoffs inference in high-dimensional models. The recipe of the method is constructing the knockoff variables by assuming a latent factor model that is exploited widely in economics and finance for the association structure of covariates. Our method and work are distinct from the existing literature in that we estimate the covariate distribution from data instead of assuming that it is known when constructing the knockoff variables, our procedure does not require any sample splitting, we provide theoretical justifications on the asymptotic false discovery rate control, and the theory for the power analysis is also established. Several simulation examples and the

[*] Yingying Fan is Dean's Associate Professor in Business Administration, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA (E-mail: *fanyingy@marshall.usc.edu*). Jinchi Lv is Professor and McAlister Associate Professor in Business Administration, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA (E-mail: *jinchilv@marshall.usc.edu*). Mahrad Sharifvaghefi is Ph.D. candidate, Department of Economics, University of Southern California, Los Angeles, CA 90089, USA (E-mail: *sharifva@usc.edu*). Yoshimasa Uematsu is Assistant Professor, Department of Economics and Management, Tohoku University, Sendai 980-8576, Japan (E-mail: *yoshimasa.uematsu.e7@tohoku.ac.jp*). Most of this work was completed while Uematsu visited USC Marshall as JSPS Overseas Research Fellow and Postdoctoral Scholar. This work was supported by NIH Grant 1R01GM131407-01, NSF CAREER Award DMS-1150318, a grant from the Simons Foundation, Adobe Data Science Research Award, and a Grant-in-Aid for JSPS Overseas Research Fellowship 29-60. We are grateful for comments from participants of seminars at USC, the 2018 Institute of Mathematical Statistics Asia Pacific Rim Meeting in Singapore, the 2018 International Symposium on Financial Engineering and Risk Management in Shanghai, and the 2018 ICSA China Statistics Conference in Qingdao.

real data analysis further demonstrate that the newly suggested method has appealing finite-sample performance with desired interpretability and stability compared to some popularly used forecasting methods.

Running title: IPAD

Key words: Reproducibility; Power; Big data; Interpretable forecasting; Stability; Latent factors; Model-X knockoffs; Large-scale inference and FDR; Scalability; Intertwined probabilistic factors decoupling; Lasso and random forest

1 Introduction

Forecasting is a fundamental problem that arises in economics and finance. With the availability of big data, many machine learning algorithms such as the Lasso and random forest can be resorted to for such a purpose by exploring a large pool of potential features. Many of these existing procedures provide a certain measure of feature importance which can then be utilized to judge the relative importance of selected features for the goal of interpretability. Yet the issue of stability in the sense of controlling the fraction of wrongly discovered features is still largely underdeveloped in the econometric settings. As argued in [20], it is difficult to obtain interpretability and stability simultaneously even in simple Lasso forecasting. A natural question is how to ensure both interpretability and stability for flexible forecasting.

Naturally stability is related to statistical inference. The recent years have witnessed a growing body of work on high-dimensional inference in the econometrics and statistics literature. For example, [42] proposed a simple procedure for inference of the average partial effects based on a debiased ℓ_1-regularized method in approximately sparse panel probit models. [38] used the de-sparsified estimator for constructing pointwise and group confidence sets. [43] conducted simultaneous inference for high-dimensional sparse linear models based on a bootstrap and desparsifying Lasso estimator. They also applied their procedure for the family-wise error rate control. [16] provided a double/debiased machine learning (DML) method for estimation and inference of treatment effects, which utilizes the Neyman orthogonal scores and cross-fitting. [18] then extended this idea to linear functionals. [17] considered debiased simultaneous inference in a system of high-dimensional regression equations with temporal and cross-sectional dependency based on a uniform robust post-selection procedure. [36] proposed Lasso residual-based tests for checking goodness-of-fit in (low- and) high-dimensional linear models. [29] presented a method for estimating the effect of the treatment on the outcome by using instrumental variables where the instruments are not necessarily valid.

Most existing work on high-dimensional inference for interpretable models has focused primarily on the aspects of post-selection inference known as selective inference and debiasing for regularization and machine learning methods. In real applications, one is often interested in conducting *global* inference relative to the full model as opposed to *local* inference conditional on the selected model. Moreover, many statistical inferences are based on p-values form significance testing. However, oftentimes obtaining valid p-values even for the Lasso in relatively complicated high-dimensional nonlinear models also remains largely unresolved, not to mention for the case of more complicated model fitting procedures such as random forest. Indeed high-dimensional inference is intrinsically challenging even in the parametric settings [27].

The desired property of stability for interpretable forecasting in this paper concentrates on *global* inference by controlling precisely the fraction of wrongly discovered features in high-dimensional models, which is also known as reproducible large-scale inference. Such a problem involves testing the joint significance of a large number of features simultaneously, which is known widely as the problem of multiple testing in statistical inference. For this problem, the null hypothesis for each feature states that the feature is unimportant in the joint model which can be understood as the property that this individual feature and the response are *independent* conditional on all the remaining features, while the corresponding alternative hypothesis states the opposite. Conventionally p-values from the hypothesis testing are used to decide whether or not to reject each null hypothesis with a significance level to control the probability of false discovery in a single hypothesis test, meaning rejecting the null hypothesis when it is true. When performing multiple hyothesis tests, the probability of making at least one false discovery which is known as the family-wise error rate can be inflated compared to that for the case of a single hypothesis test. The work on controlling such an error rate for multiple testing dates back to [13], where a simple, useful idea is lowering the significance level for each individual test as the target level divided by the total number of tests to be performed. The Bonferroni correction procedure is, however, well known to be conservative with relatively low power. Later on, [30] proposed a step-down procedure which is less conservative than the Bonferroni procedure. More recently, [35] suggested a procedure in which the critical values of individual tests are constructed sequentially.

A more powerful and extremely popular approach to multiple testing is the Benjamini–Hochberg (BH) procedure for controlling the false discovery rate (FDR) which was originated in [9], where the FDR is defined as the expectation of the fraction of falsely rejected null hypotheses known as the false discovery proportion. Given the p-values from the multiple hypothesis tests, this procedure sorts the p-values from low to high and chooses a simple, intuitive cutoff point, which can be viewed as an adaptive extension of the Bonferroni correction for multiple comparisons, of the p-values for rejecting the null hypotheses. The BH procedure was shown to be capable of controlling the FDR at the desired level for independent test statistics in [9] and for positive regression dependency among the test statistics in [10], where it was shown that a simple modification of the procedure can control the FDR under other forms of dependency but such a modification is generally conservative. There is a huge literature on the theory, applications, and various extensions of the original BH procedure for FDR control. See, for instance, [8] for a review of related developments, [24] for a factor model approach to FDR control under arbitrary covariance dependence, and [7] for a review of key results on estimation and inference including multiple testing with FDR control in high-dimensional models.

It is worth mentioning that [19] recently introduced a one covariate a time, multiple testing procedure for high-dimensional variable selection in linear regression models. In particular, their method was shown to have asymptotic FDR equal to the ratio of the number of pseudo signals and the total number of pseudo signals and true signals, where the true signals have nonzero regression coefficients and the pseudo signals have zero regression coefficients but nonzero marginal correlations with the response. Unlike [19], the main interest of our paper is the FDR control with respect to only the set of true signals.

The aforementioned econometric and statistical inference methods including the BH-type procedures for FDR control are all rooted on the availability and validity of computable p-values for evaluating variable importance. As mentioned before, such a prerequisite can become a luxury that is largely unclear how to obtain in high dimensions even for the case of Lasso in general nonlinear models and random forest. In contrast, [4] proposed a novel procedure named the knockoff filter for FDR control that bypasses the use of p-values in Gaussian linear model with deterministic design matrix, where the dimensionality is no larger than the sample size, and [5] generalized the method to high-dimensional linear models as a two-step procedure based on sample splitting, where a feature screening approach is used to reduce the dimensionality to below sample size (see, e.g., [23] and [25]) and then the knockoff filter is applied to the set of selected features after the screening step for selective inference. The key ingredient of the knockoff filter is constructing the so-called knockoff variables in a geometrical way that mimic perfectly the correlation structure among the original covariates and can be used as control variables to evaluate the importance of original variables. Recently, [15] extended the work of [4] by introducing the framework of model-X knockoffs for FDR control in general high-dimensional nonlinear models. A crucial distinction is that the knockoff variables are constructed in a probabilistic fashion such that the joint dependency structure of the original variables and their knockoff copies is invariant to the swapping of any set of original variables and their knockoff counterparts, which enables us to go beyond linear models and handle high dimensionality. As a result, model-X knockoffs enjoys exact finite-sample FDR control at the target level. However, a major assumption in [15] is that the joint distribution of all the covariates needs to be *known* for the valid FDR control.

Motivated by applications in economics and finance, in this paper we model the association structure of the covariates using the latent factor model, which reduces effectively the dimensionality and enables reliable estimation of the *unknown* joint distribution of all the covariates. By taking into account the latent factor model structure, we first estimate the association structure of covariates and then construct *empirical* knockoffs matrix using the estimated dependency structure. Our empirical knockoffs matrix can be regarded as an approximation to the *oracle* knockoffs matrix in [15] that requires the knowledge of the true covariate distribution. Exploiting the general framework of model-X knockoffs in [15], we suggest the new method of intertwined probabilistic factors decoupling (IPAD) for stable interpretable forecasting with knockoffs inference in high-dimensional models. The innovations of our method and work are fourfold. First, we estimate the covariate distribution from data instead of assuming that it is known when constructing the knockoff variables. Second, our procedure does not require any sample splitting and is thus more practical when the sample size is limited. Third, we provide theoretical justifications on the asymptotic false discovery rate control when the estimated dependency structure is employed. Fourth, the theory for power analysis is also established which reveals that there can be asymptotically no power loss in applying the knockoffs procedure compared to the underlying variable selection method. Therefore, FDR control by knockoffs can be a pure gain. Compared to earlier work, an additional challenge of our study is that knowing the true underlying distribution does *not* lead to the most efficient construction of the oracle knockoffs matrix due to the presence of latent factors. The appealing interpretability and stability of our new method compared to some popularly used forecasting methods are confirmed with several simulation and real data examples.

The rest of the paper is organized as follows. Section 2 introduces the model setting with a review of the model-X knockoffs inference framework and presents the new IPAD procedure. We establish the asymptotic properties of IPAD in Section 3. Sections 4 and 5 present several simulation and real data examples to showcase the finite-sample performance and the advantages of our newly suggested procedure compared to some popularly used ones. We discuss some implications and extensions of our work in Section 6. The proofs of the main results and additional technical details are relegated to the Appendix.

2 Intertwined probabilistic factors decoupling

To facilitate the technical presentation, we will introduce the model setting for the high-dimensional FDR control problem in Section 2.1 with a review of the model-X knockoffs inference framework in Section 2.2, and present the new IPAD procedure in Section 2.3.

2.1 Model setting

Consider the high-dimensional linear regression model

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

where $\mathbf{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the random matrix of a large number of potential regressors, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)' \in \mathbb{R}^p$ is the regression coefficient vector, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the vector of model errors, and n and p denote the sample size and dimensionality, respectively. Here without loss of generality, we assume that both the response and the covariates are centered with mean zero and thus there is no intercept. Motivated by many applications in economics and finance, we further assume that the design matrix $\mathbf{X}$ follows the *exact* factor model

$$\mathbf{X} = \mathbf{F}^0\boldsymbol{\Lambda}^{0\prime} + \mathbf{E} = \mathbf{C}^0 + \mathbf{E}, \tag{2}$$

where $\mathbf{F}^0 = (\mathbf{f}_1^0, \ldots, \mathbf{f}_n^0)' \in \mathbb{R}^{n\times r}$ is a random matrix of latent factors, $\boldsymbol{\Lambda}^0 = (\boldsymbol{\lambda}_1^0, \ldots, \boldsymbol{\lambda}_p^0)' \in \mathbb{R}^{p\times r}$ is a matrix of deterministic factor loadings, and $\mathbf{E} \in \mathbb{R}^{n\times p}$ captures the remaining variation that cannot be explained by these latent factors. We assume that the number of factors r is fixed but *unknown* and the components of $\mathbf{E}$ are independent and identically distributed (i.i.d.) from some unknown parametric distribution with cumulative distribution function $G(\cdot;\boldsymbol{\eta}^0)$, where $\boldsymbol{\eta}^0 \in \mathbb{R}^m$ is a finite-dimensional parameter vector. For simplicity, models (1) and (2) are assumed to have no endogeneity.

In this paper, we focus on the high-dimensional scenario when the dimensionality p can be much larger than sample size n. Therefore, to ensure model identifiability we impose the sparsity assumption that the true regression coefficient vector $\boldsymbol{\beta}$ has only a small portion of nonzeros; specifically, $\boldsymbol{\beta}$ takes nonzero values only on some (unknown) index set $\mathcal{S}^0 \subset \{1,\dots,p\}$ and $\beta_j = 0$ for all $j \in \mathcal{S}^1 := \{1,\dots,p\}\backslash\mathcal{S}^0$. Denote by $s = |\mathcal{S}^0|$ the size of $\mathcal{S}^0$. We assume that $s = o(n)$ throughout the paper.

We are interested in identifying the index set $\mathcal{S}^0$ with a theoretically guaranteed error rate. To be more precise, we try to select variables in $\mathcal{S}^0$ while keeping the false discovery rate (FDR) under some prespecified desired level $q \in (0,1)$, where the FDR is defined as

$$\mathrm{FDR} := \mathbb{E}\left[\mathrm{FDP}\right] \quad \text{with} \quad \mathrm{FDP} := \frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^1|}{|\widehat{\mathcal{S}}| \vee 1}. \tag{3}$$

Here the FDP stands for the false discovery proportion and $\widehat{\mathcal{S}}$ represents the set of variables selected by some procedure using observed data $(\mathbf{X},\mathbf{y})$. A slightly modified version of FDR is defined as

$$\mathrm{mFDR} := \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^1|}{|\widehat{\mathcal{S}}| + q^{-1}}\right]. \tag{4}$$

Clearly, FDR is more conservative than mFDR in that the latter is always under control if the former is.

It is easy to see that FDR is a measurement of type I error for variable selection. The other important aspect of variable selection is power, which is defined as

$$\text{Power} := \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^0|}{|\mathcal{S}^0|}\right] = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^0|}{s}\right]. \tag{5}$$

It is well known that FDR and power are two sides of the same coin. We aim at developing a variable selection procedure with theoretically guaranteed FDR control and meanwhile achieving high power.

2.2 Review of model-X knockoffs framework

The key idea of the model-X knockoffs framework is to construct the so-called model-X knockoff variables, which were introduced originally in [15] and whose definition is stated formally as follows for completeness.

Definition 1 (Model-X knockoff variables [15]) For a set of random variables $\mathbf{x} = (X_1,\dots,X_p)$, a new set of random variables $\widetilde{\mathbf{x}} = (\widetilde{X}_1,\cdots,\widetilde{X}_p)$ is called a set of model-X knockoff variables if it satisfies the following properties:

1) For any subset $\mathcal{S} \subset \{1,\dots,p\}$, we have $[\mathbf{x},\widetilde{\mathbf{x}}]_{\mathsf{swap}(\mathcal{S})} = [\mathbf{x},\widetilde{\mathbf{x}}]$ in distribution, where the vector $[\mathbf{x},\widetilde{\mathbf{x}}]_{\mathsf{swap}(\mathcal{S})}$ is obtained by swapping X_j and $\widetilde{X}_j$ for each $j \in \mathcal{S}$.

2) Conditional on $\mathbf{x}$, the knockoffs vector $\widetilde{\mathbf{x}}$ is independent of response Y.

An important consequence is that the null regressors $\{X_j : j \in \mathcal{S}^1\}$ can be swapped with their knockoffs without changing the joint distribution of the original variables $\mathbf{x}$, their knockoffs $\widetilde{\mathbf{x}}$, and response Y. That is, we can obtain for any $\mathcal{S} \subset \mathcal{S}^1$,

$$([\mathbf{x}, \widetilde{\mathbf{x}}]_{\mathsf{swap}(\mathcal{S})}, Y) \overset{d}{=} ([\mathbf{x}, \widetilde{\mathbf{x}}], Y), \tag{6}$$

where $\overset{d}{=}$ denotes equal in distribution. Such a property is known as the *exchangeability property* using the terminology in [15]. For more details, see Lemma 3.2 therein. Following [15], one can obtain a knockoffs matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ given observed design matrix $\mathbf{X}$.

Using the augmented design matrix $[\mathbf{X}, \widetilde{\mathbf{X}}]$ and response vector $\mathbf{y}$ constructed by stacking the n observations, [15] suggested constructing knockoff statistics $W_j = w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$, $j \in \{1, \dots, p\}$, for measuring the importance of the jth variable, where w_j is some function that satisfies the property that swapping $\mathbf{x}_j \in \mathbb{R}^n$ with its corresponding knockoff variable $\widetilde{\mathbf{x}}_j \in \mathbb{R}^n$ changes the sign of W_j; that is,

$$w_j([\mathbf{X}, \widetilde{\mathbf{X}}]_{\mathsf{swap}(\mathcal{S})}, \mathbf{y}) = \begin{cases} w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}), & j \notin \mathcal{S}, \\ -w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}), & j \in \mathcal{S}. \end{cases} \tag{7}$$

The knockoff statistics constructed above $W_j = w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$ satisfy the so-called sign-flip property; that is, conditional on $|W_j|$'s the signs of the null W_j's with $j \notin \mathcal{S}^0$ are i.i.d. coin flips (with equal chance 1/2). For the examples on valid constructions of knockoff statistics, see [15].

Let $t > 0$ be a fixed threshold and define $\widehat{\mathcal{S}} = \{j : W_j \geq t\}$ as the set of discovered variables. Then intuitively, the sign-flip property entails

$$\left|\widehat{\mathcal{S}} \cap \mathcal{S}^1\right| \overset{d}{=} \left|\{j : W_j \leq -t\} \cap \mathcal{S}^1\right| \leq \left|\{j : W_j \leq -t\}\right|.$$

Therefore, the FDP function can be estimated (conservatively) as

$$\mathrm{FDP} = \frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^1|}{|\widehat{\mathcal{S}}| \vee 1} \leq \frac{|\{j : W_j \leq -t\}|}{|\widehat{\mathcal{S}}| \vee 1} =: \widehat{\mathrm{FDP}}$$

for each t. In light of this observation, [15] proposed to choose the threshold by resorting to the above $\widehat{\mathrm{FDP}}$. Their results are summarized formally as follows.

Result 1 ([15]) *Let $q \in (0,1)$ denote the target FDR level. Assume that we choose a threshold $T_1 > 0$ such that*

$$T_1 = \min\left\{t > 0 : \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q\right\}$$

or $T_1 = +\infty$ if the set is empty. Then the procedure selecting the variables $\widehat{\mathcal{S}} = \{j : W_j \geq T_1\}$ controls the mFDR in (4) to no larger than q. Moreover, assume that we choose a slightly more conservative threshold $T_2 > 0$ such that

$$T_2 = \min\left\{t > 0 : \frac{1 + |\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q\right\}$$

or $T_2 = +\infty$ if the set is empty. Then the procedure selecting the variables $\widehat{\mathcal{S}} = \{j : W_j \geq T_2\}$ controls the FDR in (3) to no larger than q.

It is worth noting that Result 1 was derived under the assumption that the joint distribution of the p covariates is known. In our model setting (1) and (2), however there exist unknown parameters that need to be estimated from data. In such case, it is natural to construct the knockoff variables and knockoff statistics with estimated distribution of the p covariates. Such a plug-in principle usually leads to breakdown of the exchangeability property in Definition 1, preventing us from using directly Result 1. To address this challenging issue, we will introduce our new method in the next section and provide detailed theoretical analysis for it.

It is also worth mentioning that recently, [6] provided an elegant new line of theory which ensures FDR control of model-X knockoffs procedure under the approximate exchangeability assumption, which is weaker than the exact exchangeability condition required in Definition 1. However, the conditions they need on estimation error of the joint distribution of $\mathbf{x}$ is difficult to be satisfied in high dimensions. [26] investigated the robustness of model-X knockoffs procedure with respect to unknown covariate distribution when covariates $\mathbf{x}$ follow a joint Gaussian distribution. Their procedure needs data splitting and their proofs rely heavily on the Gaussian distribution assumption, and thus their development may not be suitable for economic data with limited sample size and heavy-tailed distribution. For these reasons, our results complement substantially those in [15], [26], and [6].

2.3 IPAD

It has been seen from the previous section that the key for the model-X knockoffs framework is the construction of valid knockoff variables. We begin with the ideal situation where the the factor model structure (2) is fully available to us; that is, we know the realization $\mathbf{C}^0$ and the distribution $G(\cdot; \boldsymbol{\eta}^0)$ for the error matrix $\mathbf{E}$. In such case, the oracle knockoffs matrix $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ can be constructed as

$$\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0) = \mathbf{C}^0 + \mathbf{E}_{\boldsymbol{\eta}^0}, \tag{8}$$

where $\mathbf{E}_{\boldsymbol{\eta}^0}$ is an i.i.d. copy of $\mathbf{E}$ and $\boldsymbol{\theta}^0 = (\mathbf{C}^0, \boldsymbol{\eta}^0)$ is the augmented parameter vector. Note that $\mathbf{E}_{\boldsymbol{\eta}_0}$ itself is not a function of $\boldsymbol{\eta}_0$, but we slightly abuse the notation to emphasize the dependence of the distribution function on parameter $\boldsymbol{\eta}_0$. It is easy to check that $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ constructed above is a valid knockoffs matrix and satisfies the properties in Definition 1. Although $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ is generally unavailable to us, it plays an important role in our theoretical developments.

We remark that in the construction above, we slightly misuse the concept and call $\mathbf{C}^0$ a parameter. This is because although $\mathbf{C}^0$ is a random matrix, for the construction of valid knockoff variables it is the particular realization $\mathbf{C}^0$ leading to the observed data matrix $\mathbf{X}$ that matters. In other words, a valid construction of knockoff variables requires the knowledge of the specific realization $\mathbf{C}^0$ instead of the distribution of $\mathbf{C}^0$. To understand this, consider the scenario where the underlying parameter $\boldsymbol{\eta}^0$ and the exact distribution of

$\mathbf{C}^0$ are fully available to us. If we independently generate random variables from this known distribution and form a new data matrix $\mathbf{X}_1$, because of the independence between $\mathbf{X}_1$ and $\mathbf{X}$, the exchangeability assumption in Definition 1 will be violated and thus $\mathbf{X}_1$ cannot be a valid knockoffs matrix. On the other hand, as long as we know the realization $\mathbf{C}^0$ and parameter $\boldsymbol{\eta}^0$, a valid knockoffs matrix $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ can be constructed using (8) regardless of whether the exact distribution of $\mathbf{C}^0$ is available to us or not.

In practice, however $\boldsymbol{\theta}^0$ is unavailable to us and consequently, $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ is inaccessible. To overcome this difficulty, we next introduce our new method IPAD. We start with introducing the *knockoff generating function* – for each given parameter vector $\boldsymbol{\theta} = (\mathbf{C}, \boldsymbol{\eta})$, define

$$\widetilde{\mathbf{X}}(\boldsymbol{\theta}) = \mathbf{C} + \mathbf{E}_{\boldsymbol{\eta}}, \tag{9}$$

where $\mathbf{E}_{\boldsymbol{\eta}}$ is a matrix composed of i.i.d. random samples from the distribution $G(\cdot; \boldsymbol{\eta})$. Letting $\hat{\boldsymbol{\theta}}$ denote an estimator (obtained using data $\mathbf{X}$) of $\boldsymbol{\theta}^0$, we name $\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})$ as the *empirical* knockoffs matrix while $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ as the *oracle (ideal)* knockoffs matrix.

With the aid of empirical knockoffs matrix, we suggest the following IPAD procedure for FDR control with knockoffs inference.

Procedure 1 (IPAD) 1) (Estimation of parameters) Estimate the unknown parameters in $\boldsymbol{\theta}^0$ using the design matrix $\mathbf{X}$. Denote by $\hat{\boldsymbol{\theta}} = (\widehat{\mathbf{C}}, \hat{\boldsymbol{\eta}})$ the resulting estimated parameter vector.

2) (Construction of empirical knockoffs matrix) Construct the empirical knockoffs matrix by applying the knockoff generating function in (9) to the estimated parameter $\hat{\boldsymbol{\theta}}$; that is,

$$\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}) = \widehat{\mathbf{C}} + \mathbf{E}_{\hat{\boldsymbol{\eta}}}, \tag{10}$$

where $\mathbf{E}_{\hat{\boldsymbol{\eta}}} \in \mathbb{R}^{n \times p}$ is a matrix composed of i.i.d. random variables from $G(\cdot; \hat{\boldsymbol{\eta}})$, and is independent of $(\mathbf{X}, \mathbf{y})$ conditional on $\hat{\boldsymbol{\eta}}$.

3) (Application of knockoffs inference) Calculate knockoff statistics $W_j(\hat{\boldsymbol{\theta}})$ using data $([\mathbf{X}, \widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})], \mathbf{y})$ and then construct $\widehat{\mathcal{S}}$ by applying knockoffs inference to $W_j(\hat{\boldsymbol{\theta}})$.

Intuitively, the accuracy of the estimator $\hat{\boldsymbol{\theta}}$ in Step 1 will affect the performance of our IPAD procedure. In fact, as shown later in our Theorem 1, the consistency rate of $\hat{\boldsymbol{\theta}}$ is indeed reflected in the asymptotic FDR control of the IPAD procedure. For the specific case when the error distribution is $N(0, \sigma^2)$, the parameter σ^2 can be estimated naturally as $(np)^{-1} \sum_{i,j} \hat{e}_{ij}^2$, where $\hat{e}_{ij}$'s are the entries of $\widehat{\mathbf{E}} = \mathbf{X} - \widehat{\mathbf{C}}$. In Step 3, various methods can be used to construct knockoff statistics. For the illustration purpose, we use the *Lasso coefficient difference* (LCD) statistic as in [15]. Specifically, with $\mathbf{y}$ the response vector and $([\mathbf{X}, \widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})])$ the augmented design matrix we consider the variable selection procedure Lasso [39] which solves the following optimization problem

$$\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda) = \arg \min_{\mathbf{b} \in \mathbb{R}^{2p}} \left\{ \|\mathbf{y} - [\mathbf{X}, \widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})]\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}, \tag{11}$$

where $\lambda \geq 0$ is the regularization parameter and $\|\cdot\|_m$ with $m \geq 1$ denotes the vector ℓ_m-norm. Then for each variable $\mathbf{x}_j$, the knockoff statistic can be constructed as

$$W_j(\hat{\boldsymbol{\theta}}; \lambda) = |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)| - |\hat{\beta}_{p+j}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)|, \tag{12}$$

where $\hat{\beta}_\ell^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)$ is the ℓth component of the Lasso regression coefficient vector $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)$. It is seen that intuitively the LCD knockoff statistics evaluate the relative importance of the jth original variable by comparing its Lasso coefficient $\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)$ with that of its knockoff copy $\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}; \lambda)$. In the ideal case when the oracle knockoffs matrix $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ is used instead of $\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})$ in (11), it is easy to verify that the LCD is a valid construction of knockoff statistics and satisfies the sign-flip property in (7). Consequently, the general theory in [15] can be applied to show that the FDR is controlled in finite sample. We next show that even with the empirical knockoffs matrix employed in (11), the FDR can still be asymptotically controlled with delicate technical analyses.

3 Asymptotic properties of IPAD

We now provide theoretical justifications for our IPAD procedure suggested in Section 2 with the LCD knockoff statistics $W_j(\hat{\boldsymbol{\theta}}; \lambda) = w_j([\mathbf{X}, \widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})], \mathbf{y}; \lambda)$ defined in (12). We will first present some technical conditions in Section 3.1, then prove in Section 3.2 that the FDR is asymptotically under control at desired target level q, and finally in Section 3.3 show that asymptotically IPAD has no power loss compared to the Lasso under some regularity conditions.

3.1 Technical conditions

We first introduce some notation and definitions which will be used later on. We use $X \sim \mathrm{subG}(C_x^2)$ to denote that X is a sub-Gaussian random variable with *variance proxy* $C_x^2 > 0$ if $\mathbb{E}[X] = 0$ and its tail probability satisfies $\mathbb{P}(|X| > u) \leq 2\exp(u^2/C_x^2)$ for each $u \geq 0$. In all technical assumptions below, we use $M > 1$ to denote a large enough generic constant. Throughout the paper, for any vector $\mathbf{v} = (v_i)$ let us denote by $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_{\max}$ the ℓ_1-norm, ℓ_2-norm, and max-norm defined as $\|\mathbf{v}\|_1 = \sum_i |v_i|$, $\|\mathbf{v}\|_2 = (\sum_i v_i^2)^{1/2}$, and $\|\mathbf{v}\|_{\max} = \max_i |v_i|$, respectively. For any matrix $\mathbf{M} = (m_{ij})$, we denote by $\|\mathbf{M}\|_F$, $\|\mathbf{M}\|_1$, $\|\mathbf{M}\|_2$, and $\|\mathbf{M}\|_{\max}$ the Frobenius norm, entrywise ℓ_1-norm, spectral norm, and entrywise ℓ_∞-norm defined as $\|\mathbf{M}\|_F = \|\operatorname{vec}(\mathbf{M})\|_2$, $\|\mathbf{M}\|_1 = \|\operatorname{vec}(\mathbf{M})\|_1$, $\|\mathbf{M}\|_2 = \sup_{\mathbf{v}\neq\mathbf{0}} \|\mathbf{M}\mathbf{v}\|_2/\|\mathbf{v}\|_2$, and $\|\mathbf{M}\|_{\max} = \|\operatorname{vec}(\mathbf{M})\|_{\max}$, respectively, where $\operatorname{vec}(\mathbf{M})$ represents the vectorization of matrix $\mathbf{M}$. For a symmetric matrix $\mathbf{M}$, $\operatorname{vech}(\mathbf{M})$ stands for the vectorization of the lower triangular part of $\mathbf{M}$.

Condition 1 (Regression errors) The model error vector $\boldsymbol{\varepsilon}$ has i.i.d. components from $\mathrm{subG}(C_\varepsilon^2)$.

Condition 2 (Latent factors) The rows of $\mathbf{F}^0$ consist of mean zero i.i.d. random vectors $\mathbf{f}_i^0 \in \mathbb{R}^r$ such that $\|\mathbf{F}^0\|_{\max} \leq M$ almost surely (a.s.) and $\|\boldsymbol{\Sigma}_f\|_2 + \|\boldsymbol{\Sigma}_f^{-1}\|_2 \leq M$, where $\boldsymbol{\Sigma}_f := \mathbb{E}[\mathbf{f}_i^0\mathbf{f}_i^{0\prime}]$.

Condition 3 (Factor loadings) The rows of $\mathbf{\Lambda}^0$ consist of deterministic vectors $\boldsymbol{\lambda}_j^0 \in \mathbb{R}^r$ such that $\|\mathbf{\Lambda}^0\|_{\max} \leq M$ and $\|p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\|_2 + \|(p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0)^{-1}\|_2 \leq M$.

Condition 4 (Factor errors) The entries of matrix $\mathbf{E}_{\boldsymbol{\eta}^0}$ are i.i.d. copies of $e_{\boldsymbol{\eta}^0} \sim \mathrm{subG}(C_e^2)$ with continuous distribution function $G(\cdot;\boldsymbol{\eta}^0)$. For each $1 \leq \ell \leq m$, the ℓth element of $\boldsymbol{\eta}^0$ is specified as $\eta_\ell^0 = h_\ell(\mathbb{E}[e_{\boldsymbol{\eta}^0}], \ldots, \mathbb{E}[e_{\boldsymbol{\eta}^0}^m])$ with $h_\ell : \mathbb{R}^m \to \mathbb{R}$ some local Lipschitz continuous function in the sense that

$$\Big|h_\ell(t_1, \ldots, t_m) - h_\ell(\mathbb{E}[e_{\boldsymbol{\eta}^0}], \ldots, \mathbb{E}[e_{\boldsymbol{\eta}^0}^m])\Big| \leq M \max_{k\in\{1,\ldots,m\}} \Big|t_k - \mathbb{E}[e_{\boldsymbol{\eta}^0}^k]\Big|$$

for each $t_k \in \{t : |t - \mathbb{E}[e_{\boldsymbol{\eta}^0}^k]| \leq Mc_{np}\}$ and $1 \leq k \leq m$, where $c_{np} := (p^{-1}\log n)^{1/2} + (n^{-1}\log p)^{1/2}$. Moreover, there exists some stochastic process $(e_{\boldsymbol{\eta}})_{\boldsymbol{\eta}}$ such that

i) for each $\boldsymbol{\eta} \in \{\boldsymbol{\eta} \in \mathbb{R}^m : \|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|_{\max} \leq Mc_{np}\}$, the entries of $\mathbf{E}_{\boldsymbol{\eta}}$ in (9) have identical distribution to $e_{\boldsymbol{\eta}}$,

ii) for some sub-Gaussian random variable $Z \sim \mathrm{subG}\left(c_e^2\right)$ with some positive constant c_e,

$$\sup_{\boldsymbol{\eta}:\, \|\boldsymbol{\eta}-\boldsymbol{\eta}^0\|_{\max}\leq Mc_{np}} |e_{\boldsymbol{\eta}} - e_{\boldsymbol{\eta}^0}| \leq M^{1/2}c_{np}^{1/2}|Z|. \tag{13}$$

Condition 5 (Eigenseparation) The r eigenvalues of $p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{\Sigma}_f$ are distinct for all p.

The number of factors r is assumed to be known for developing the theory with simplification, but in practice it can be estimated consistently using methods such as information criteria [3] and test statistics [1]. The sub-Gaussian assumptions in Conditions 1 and 4 can be replaced with some other tail conditions as long as similar concentration inequalities hold. Condition 3 is standard in the analysis of factor models. Stochastic loadings can be assumed in Condition 3 with some appropriate distributional assumption, such as sub-Gaussianity, at the cost of much more tedious technical arguments. The boundedness of the eigenvalues of $\mathbf{\Sigma}_f$ in Condition 2 is standard while the i.i.d. assumption and boundedness of $\mathbf{f}_i^0$ are stronger compared to the existing literature (e.g., [3] and [2]). However, these conditions are imposed mostly for technical simplicity. In fact, the boundedness condition on $\mathbf{f}_i^0$ can be replaced with (unbounded) sub-Gaussian or other heavier-tail assumption whenever concentration inequalities are available at the cost of slower convergence rates and stronger sample size requirement. Our theory on FDR control is based on that in [15], which applies only to the case of i.i.d. rows of design matrix $\mathbf{X}$. This is the main reason for imposing the i.i.d. assumption on ε_i and $\mathbf{f}_i$ in Conditions 1 and 2. However, we conjecture that similar results can also hold in the presence of some sufficiently weak serial dependence in ε_i and $\mathbf{f}_i$. Condition 4 introduces a *sub-Gaussian process* $e_{\boldsymbol{\eta}}$ with respect to $\boldsymbol{\eta}$. The norm in (13) can be replaced with any other norm since $\boldsymbol{\eta}$ is finite dimensional. In the specific case when the components of $\mathbf{E}$ have Gaussian distribution such that $\boldsymbol{\eta}$ is a scalar parameter representing variance, by the the reflection principle for the Wiener process ([12], p.511), $e_{\boldsymbol{\eta}}$ can be constructed as a Wiener process and the inequality (13) can be satisfied. For more information on sub-Gaussian processes, see, e.g., [41]. Condition 5 guarantees that $\widehat{\mathbf{F}}'\mathbf{F}^0/n$ is asymptotically nonsingular, which has been proved in [2] and is used in the proof of Lemma 6 in Appendix.

Recall that in the IPAD procedure, we first obtain the augmented Lasso estimator $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda) \in \mathbb{R}^{2p}$ by regressing $\mathbf{y}$ on $[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]$. Denote by $\mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda) = \mathrm{supp}(\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda)) \subset \{1,\ldots,2p\}$ the active set of the augmented Lasso regression coefficient vector. Throughout this section, we content ourselves with sparse estimates satisfying

$$|\mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda)| \leq k/2 \tag{14}$$

for some positive integer k which may diverge with n at an order to be specified later; see, e.g., [28] and [32] for a similar constraint and justifications therein. This can always be achieved since users have the freedom to choose the size of the Lasso model.

3.2 FDR control

To develop the theory for IPAD, we consider the principle component estimator $\widehat{\mathbf{C}}$ for the realization $\mathbf{C}^0$. More specifically, we first conduct the singular value decomposition (SVD) $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$ with $\mathbf{U}$ and $\mathbf{V}$ the left and right singular matrices and $\mathbf{D}$ a diagonal matrix of singular values, and then threshold the diagonal matrix $\mathbf{D}$ by setting the smallest $n-r$ singular values to zero. Let us denote the thresholded matrix as $\mathbf{D}_r$. Then matrix $\mathbf{C}^0$ can be estimated as $\widehat{\mathbf{C}} = \mathbf{U}\mathbf{D}_r\mathbf{V}'$. Denote by $\widehat{\mathbf{E}} = (\hat{e}_{ij}) = \mathbf{X} - \widehat{\mathbf{C}}$. The estimator $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \cdots, \hat{\eta}_m)'$ is constructed as $\hat{\eta}_\ell = h_\ell(\mathbb{E}_{np}\,\hat{e}, \ldots, \mathbb{E}_{np}\,\hat{e}^m)$ with h_ℓ, $1 \leq \ell \leq m$, introduced in Condition 4 and $\mathbb{E}_{np}\hat{e}^k = (np)^{-1}\sum_{i,j}\hat{e}_{ij}^k$ the empirical moments of $\hat{e}_{ij}$. Throughout our theoretical analysis, we consider the regularization parameter fixed at $\lambda = C_0 n^{-1/2}\log p$ with C_0 some large enough constant for all the Lasso procedures. Therefore, we will drop the dependence of various quantities on λ whenever there is no confusion. For example, we will write $\mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda)$ and $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta};\lambda)$ as $\mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$, respectively.

Denote by $\mathbf{U}(\boldsymbol{\theta}) := n^{-1}[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]'[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]$ and $\mathbf{v}(\boldsymbol{\theta}) := n^{-1}[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]'\mathbf{y}$ and define $\mathbf{T}(\boldsymbol{\theta}) := \mathrm{vec}(\mathrm{vech}\,\mathbf{U}(\boldsymbol{\theta}), \mathbf{v}(\boldsymbol{\theta})) \in \mathbb{R}^P$ with $P := p(2p+3)$. The following lemma states that the statistic $\mathbf{T}(\boldsymbol{\theta})$ plays a crucial role in our procedure.

Lemma 1 *The set of variables $\widehat{S}$ selected by Procedure 1 depends only on $\mathbf{T}(\boldsymbol{\theta})$.*

For any given $\boldsymbol{\theta}$, define the active set $\mathcal{A}^*(\boldsymbol{\theta}) := \mathcal{A}_1^{\mathsf{aug}}(\boldsymbol{\theta}) \cup \mathcal{A}_2^{\mathsf{aug}}(\boldsymbol{\theta}) \subset \{1,\ldots,p\}$, where $\mathcal{A}_1^{\mathsf{aug}}(\boldsymbol{\theta}) := \{j : j \in \{1,\ldots,p\} \cap \mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta})\}$ and $\mathcal{A}_2^{\mathsf{aug}}(\boldsymbol{\theta}) := \{j-p : j \in \{p+1,\ldots,2p\} \cap \mathcal{A}^{\mathsf{aug}}(\boldsymbol{\theta})\}$. That is, $\mathcal{A}^*(\boldsymbol{\theta})$ is equal to the support of knockoff statistics $(W_1(\boldsymbol{\theta}), \cdots, W_p(\boldsymbol{\theta}))'$ if there are no ties on the magnitudes of the augmented Lasso coefficient vector $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$.

We next focus on the low-dimensional structure of $\mathbf{T}(\boldsymbol{\theta})$ inherited from the augmented Lasso because it will be made clear that this is the key to controlling the FDR *without* sample splitting. For any subset $\mathcal{A} \subset \{1,\ldots,p\}$, define a lower-dimensional expression of the vector as $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) := \mathrm{vec}(\mathrm{vech}\,\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}), \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}))$ with $\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta})$ the principle submatrix of $\mathbf{U}(\boldsymbol{\theta})$ formed by columns and rows in $\mathcal{A}$ and $\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta})$ the subvector of $\mathbf{v}(\boldsymbol{\theta})$ formed by components in $\mathcal{A}$. Then it is easy to see that $\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})]'[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})]$ and $\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})]'\mathbf{y}$. Motivated by Lemma 1, we define a family of mappings indexed by $\mathcal{A}$ that describes the *selection algorithm* of Procedure 1 with given data set $([\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})], \mathbf{y})$ that forms $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta})$. Formally, define a mapping $S_{\mathcal{A}} : \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)} \to 2^{\mathcal{A}}$ as $\mathbf{t}_{\mathcal{A}} \mapsto S_{\mathcal{A}}(\mathbf{t}_{\mathcal{A}})$ for given $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) = \mathbf{t}_{\mathcal{A}}$, where $2^{\mathcal{A}}$ refers to the power set of $\mathcal{A}$. That is, $S_{\mathcal{A}}(\mathbf{t}_{\mathcal{A}})$ represents the outcome of first restricting ourselves to the smaller set of variables $\mathcal{A}$ and then applying IPAD to $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) = \mathbf{t}_{\mathcal{A}}$ to further select variables from set $\mathcal{A}$.

Lemma 2 *Under Conditions 1–4, for any subset $\mathcal{A} \supset \mathcal{A}^*(\boldsymbol{\theta})$ we have $S_{\{1,\ldots,p\}}(\mathbf{T}(\boldsymbol{\theta})) = S_{\mathcal{A}}(\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}))$.*

When restricting on set $\mathcal{A}$, we can apply Procedure 1 to a lower-dimensional data set $([\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})], \mathbf{y})$ that forms $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta})$ to further select variables from $\mathcal{A}$. The previous two lemmas ensure that this gives us a subset of $\mathcal{A}$ that is identical to $S_{\{1,\ldots,p\}}(\mathbf{T}(\boldsymbol{\theta}))$. Note that the lower-dimensional problem based on $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta})$ can be easier compared to the original one. We also would like to emphasize that the dimensionality reduction to a smaller model $\mathcal{A}$ is only for assisting the theoretical analysis and our Procedure 1 does not need any knowledge of such set $\mathcal{A}$.

It is convenient to define $\mathbf{t}_0 = \mathbb{E}\,\mathbf{T}(\boldsymbol{\theta}^0) \in \mathbb{R}^P$. Denote by

$$\mathbb{I} := \left\{ \mathbf{t} \in \mathbb{R}^P : \|\mathbf{t} - \mathbf{t}_0\|_{\max} \le a_{np} := C_1 (k^{1/2} + s^{3/2}) \tilde{c}_{np} \right\}, \tag{15}$$

where C_1 is some positive constant and $\tilde{c}_{np} = p^{-1/2} \log n + n^{-1/2} \log p$. For any subset $\mathcal{A} \subset \{1, \cdots, p\}$, let $\mathbb{I}_{\mathcal{A}}$ be the subspace of $\mathbb{I}$ when taking out the coordinates corresponding to $\mathbb{E}\,\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}^0)$. Thus $\mathbb{I}_{\mathcal{A}} \subset \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)}$. In addition to Conditions 1–5, we need an assumption on the algorithmic stability of Procedure 1.

Condition 6 (Algorithmic stability) For any subset $\mathcal{A} \subset \{1, \ldots, p\}$ that satisfies $|\mathcal{A}| \le k \le n \wedge p$, there exists a positive sequence $\rho_{np} \to 0$ as $n \wedge p \to \infty$ such that

$$\sup_{|\mathcal{A}| \le k} \sup_{\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{I}_{\mathcal{A}}} \frac{\left| S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1) \right|}{|S_{\mathcal{A}}(\mathbf{t}_1)| \wedge |S_{\mathcal{A}}(\mathbf{t}_2)|} = O(\rho_{np}),$$

where $\triangle$ stands for the symmetric difference between two sets.

Intuitively the above condition assumes that the knockoffs procedure is stable with respect to a small perturbation to the input $\mathbf{t}$ in any lower-dimensional subspace $\mathbb{I}_{\mathcal{A}}$. Under these regularity conditions, the asymptotic FDR control of our IPAD procedure can be established.

Theorem 1 (Robust FDR control) *Assume that Conditions 1–6 hold. Fix an arbitrary positive constant ν. If (s, k, n, p) satisfies $s \vee k \le n \wedge p$, $c_{np} \le c/[r^2 M^2 C(\nu + 2)]^{1/2}$, and $(k^{1/2} + s^{3/2}) \tilde{c}_{np} \to 0$ as $n \wedge p \to \infty$ with c and C some positive constants defined in Lemma 7 in Appendix, then the set of variables $\widehat{\mathcal{S}}$ obtained by Procedure 1 (IPAD) with the LCD knockoff statistics controls the FDR (3) to be no larger than $q + O\left(\rho_{np} + n^{-\nu} + p^{-\nu}\right)$.*

Recall that by definition, the FDR is a function of $\mathbf{T}(\hat{\boldsymbol{\theta}})$ and can be written as $\mathbb{E}\,\mathrm{FDP}(\mathbf{T}(\hat{\boldsymbol{\theta}}))$ while the FDR computed with the oracle knockoffs, $\mathbb{E}\,\mathrm{FDP}(\mathbf{T}(\boldsymbol{\theta}^0))$, is perfectly controlled to be no larger than q. This observation motivates us to first establish asymptotic equivalence of $\mathbf{T}(\hat{\boldsymbol{\theta}})$ and $\mathbf{T}(\boldsymbol{\theta}^0)$ with large probability. Then a natural idea is to show that $\mathbb{E}\,\mathrm{FDP}(\mathbf{T}(\hat{\boldsymbol{\theta}}))$

converges to $\mathbb{E}\,\mathrm{FDP}(\mathbf{T}(\boldsymbol{\theta}^0))$ in probability, which turns out to be highly nontrivial because of the discontinuity of $\mathrm{FDP}(\cdot)$ (the convergence would be straightforward via the Portmanteau lemma if $\mathrm{FDP}(\cdot)$ were continuous). Condition 6 above provides a remedy to this issue by imposing the algorithmic stability assumption.

3.3 Power analysis

We have established the asymptotic FDR control for our IPAD procedure in Section 3.2. We now look at the other side of the coin – the power (5). Recall that in IPAD, we apply the knockoffs inference procedure to the knockoff statistics LCD, which are constructed using the augmented Lasso in (11). Therefore the final set of variables selected by IPAD is a subset of variables picked by the augmented Lasso. For this reason, the power of IPAD is always upper bounded by that of Lasso. We will show in this section that there is in fact no power loss relative to the augmented Lasso in the asymptotic sense.

Condition 7 (Signal strength I) For any subset $\mathcal{A} \subset \mathcal{S}^0$ that satisfies $|\mathcal{A}|/s > 1 - \gamma$ for some $\gamma \in (0, 1]$, it holds that $\|\boldsymbol{\beta}_{\mathcal{A}}\|_1 > b_{np} s n^{-1/2} \log p$ for some positive sequence $b_{np} \to \infty$.

Condition 8 (Signal strength II) There exists some constant $C_2 \in (2(qs)^{-1}, 1)$ such that $|\mathcal{S}_2| \geq C_2 s$ with $\mathcal{S}_2 = \{j : |\beta_j| \gg (s/n)^{1/2} \log p\}$.

Condition 7 requires that the overall signal is not too weak, but is weaker than the conventional beta-min condition $\min_{j \in \mathcal{S}^0} |\beta_j| \gg n^{-1/2} \log p$. Under Condition 8, we can show that $|\widehat{\mathcal{S}}| \geq C_2 s$ with probability at least $1 - O(p^{-\nu} + n^{-\nu})$ using similar techniques to those of Lemma 6 in [26]. The intuition is that given $s \to \infty$, for a variable selection procedure to have high power it should select at least a reasonably large number of variables. The result $|\widehat{\mathcal{S}}| \geq C_2 s$ will be used to derive the asymptotic order of threshold T, which is in turn crucial to establish the theorem below on power.

Theorem 2 (Power guarantee) *Assume that Conditions 1–5 and 7–8 hold. Fix an arbitrary positive constant ν. If (s, k, n, p) satisfies $2s \leq k \leq n \wedge p$, $c_{np} \leq c/(r^2 M^2 C(\nu + 2))^{1/2}$, and $sk^{1/2}\tilde{c}_{np} \to 0$ as $n \wedge p \to \infty$ with c and C some positive constants defined in Lemma 7, then both the Lasso procedure based on $(\mathbf{X}, \mathbf{y})$ and our IPAD procedure (Procedure 1) have power bounded from below by $\gamma - o(1)$ as $n \wedge p \to \infty$. In particular, if $\gamma = 1$ IPAD has no power loss compared to Lasso asymptotically.*

4 Simulation studies

We have shown in Section 3 that IPAD can asymptotically control the FDR in high-dimensional setting and there can be no power loss in applying the procedure. We next move on to numerically investigate the finite-sample performance of IPAD using synthetic data sets. We will compare IPAD with the knockoff filter in [4] (BCKnockoff) and the high-dimensional knockoff filter in [5] (HD-BCKnockoff). In what follows, we will first explain in detail the model setups and simulation settings, then discuss the implementation of the aforementioned methods, and finally summarize the comparison results.

4.1 Simulation designs and settings

In all simulations, the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is generated from the factor model

$$\mathbf{X} = \mathbf{F}^0(\mathbf{\Lambda}^0)' + \sqrt{r\theta}\mathbf{E} = \mathbf{C}^0 + \sqrt{r\theta}\mathbf{E}, \tag{16}$$

where $\mathbf{F}^0 = (\mathbf{f}_1^0, \cdots, \mathbf{f}_n^0)' \in \mathbb{R}^{n \times r}$ is the matrix of latent factors, $\mathbf{\Lambda}^0 = (\boldsymbol{\lambda}_1^0, \ldots, \boldsymbol{\lambda}_p^0)' \in \mathbb{R}^{p \times r}$ is the matrix of factor loadings, $\mathbf{E} \in \mathbb{R}^{n \times p}$ is the matrix of model errors, and θ is a constant controlling the signal-to-noise ratio. The term $\sqrt{r}$ is used to single out the effect of the number of factors in calculating the signal-to-noise ratio in factor model (16). We then rescale each column of $\mathbf{X}$ to have ℓ_2-norm one and simulate the response vector $\mathbf{y} = (y_1, \cdots, y_n)'$ from the following model

$$y_i = f(\mathbf{x}_i) + \sqrt{c}\varepsilon_i, \; i = 1, \cdots, n, \tag{17}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is the link function which can be linear or nonlinear, $c > 0$ is a constant controlling the signal-to-noise ratio, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)'$ is the vector of model error. We next explain the four different designs of our simulation studies.

4.1.1 Design 1: linear model with normal factor design matrix

The elements of $\mathbf{F}^0$, $\mathbf{\Lambda}^0$, $\mathbf{E}$, and $\boldsymbol{\varepsilon}$ are drawn independently from $\mathcal{N}(0, 1)$. The link function takes a linear form, that is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sqrt{c}\boldsymbol{\varepsilon},$$

where the coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)' \in \mathbb{R}^p$ is generated by first choosing s random locations for the true signals and then setting β_j at each location to be either A or $-A$ randomly with A some positive value. The remaining $p - s$ components of $\boldsymbol{\beta}$ are set to zero.

4.1.2 Design 2: linear model with fat-tail factor matrix and serial dependence

The elements of $\mathbf{E}$ are generated as

$$e_{ij} = \left(\frac{\nu - 2}{\chi^2_{\nu,j}}\right) u_{ij}, \tag{18}$$

where $u_{ij} \sim i.i.d.\ \mathcal{N}(0, 1)$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$, and $\chi^2_{\nu,j}, \; j = 1, \cdots, p$ are i.i.d. random variables from chi-square distribution with $\nu = 8$ degrees of freedom. The rest of the design is the same as in Design 1. It is worth mentioning that in this case, the entries of matrix $\mathbf{E}$ have fat-tail distribution with serial dependence in each column because of the common factor $\chi^2_{\nu,j}$. This design is used to check the robustness of IPAD method with respect to the serial dependence and the fat-tail distribution of $\mathbf{E}$.

4.1.3 Design 3: linear model with misspecified design matrix

To evaluate the robustness of IPAD procedure to the misspecification of the factor model structure (16), we set $\mathbf{\Lambda} = \mathbf{0}$, $r\theta = 1$ and simulate the rows of matrix $\mathbf{E}$ independently from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\sigma_{ij})$, $\sigma_{ij} = \rho^{|i-j|}$ for $\leq i, j \leq p$. The remaining design is the same as in

Design 1. It is seen that our assumption on the independence of the entries of $\mathbf{E}$ is violated. This design is used to test the robustness of IPAD to misspecification of the factor model structure of $\mathbf{X}$.

4.1.4 Design 4: nonlinear model with normal factor design matrix

Our last design is used to evaluate the performance of IPAD method when the link function f is nonlinear. To be more specific, we assume the following nonlinear model between the response and covariates

$$\mathbf{y} = \sin(\mathbf{X}\boldsymbol{\beta}) \exp(\mathbf{X}\boldsymbol{\beta}) + \sqrt{c}\boldsymbol{\varepsilon},$$

where the coefficient vector $\boldsymbol{\beta}$, design matrix $\mathbf{X}$, and model error $\boldsymbol{\varepsilon}$ are generated similarly as in Design 1.

4.1.5 Simulation settings

The target FDR level is set to be $q = 0.2$ in all simulations. For Design 1 and Design 2, we set $n = 2000$, $p = 2000$, $A = 4$, $s = 50$, $c = 0.2$, $r = 3$, and $\theta = 1$. In order to evaluate the sensitivity of our method to the dimensionality p and the model sparsity s, we also explore the settings of $p = 1000, 3000$ and $s = 100, 150$. In Design 3, we set $r = 0$ and $\rho = 0, 0.5$. In Design 4, since the model is nonlinear, we use nonparametric method to fit the model and consider lower-dimensional settings of $p = 50, 250, 500$. We also decrease the number of observations to $n = 1000$ and number of true variables to $s = 10$. Moreover, we set $\theta = 1, 2$ and $c = 0.1, 0.2, 0.3$ to test the effects of signal-to-noise ratio on the performance of IPAD procedure in Design 4.

4.2 Estimation procedure

In implementing the IPAD algorithm suggested in Section 2, we use the PC_{p1} criterion proposed in [3] to estimate the number of factors r. With an estimated number of factors $\hat{r}$, we use the principle component method discussed in Section 3.2 to obtain an estimate $\widehat{\mathbf{C}}$ of matrix $\mathbf{C}^0$. Denote by $\widehat{\mathbf{E}} = (\hat{e}_{ij}) = \mathbf{X} - \widehat{\mathbf{C}}$. Recall that in the construction of knockoff variables, the distribution of $\mathbf{E}$ needs to be estimated. Throughout our simulation studies, we misspecify the model and treat the entries of $\mathbf{E}$ as i.i.d. Gaussian random variables. Under this working model assumption, the only unknown parameter is the variance which can be estimated by the following maximum likelihood estimator

$$\hat{\sigma}^2 = (np)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{p} \hat{e}_{ij}^2.$$

Then the knockoffs matrix $\widehat{\mathbf{X}}$ is constructed using (10) with the entries of $\mathbf{E}_{\hat{\boldsymbol{\eta}}}$ drawn independently from $\mathcal{N}(0, \hat{\sigma}^2)$. For the two comparison methods BCKnockoff and HD-BCKnockoff, we follow the implementation in [4] and [5], respectively. Thus it is seen that neither BCKnockoff nor HD-BCKnockoff uses the factor structure in $\mathbf{X}$ when constructing the knockoff variables.

In Designs 1–3, with the constructed empirical knockoffs matrix $\widehat{\mathbf{X}}$ we apply the Lasso method to fit the model with $\mathbf{y}$ the response vector and $[\mathbf{X}, \widehat{\mathbf{X}}]$ the augmented design matrix. Then the LCD discussed in Section 2.3 is used in the construction of knockoff statistics. In Design 4, we assume the nonlinear relationship between the response and the covariates. In this case, random forest is used for estimation of the model. To construct the knockoff statistics, we use the variable importance measure of mean decrease accuracy (MDA) introduced in [14]. This measure is based on the idea that if a variable is unimportant, then rearranging its values should not degrade the prediction accuracy. The MDA for the jth variable, denoted as $\widehat{\mathrm{MDA}}_j$, measures the amount of increase in prediction error when the values of the jth variable in the out-of-sample prediction are permuted randomly. Then intuitively, $\widehat{\mathrm{MDA}}_j$ will be small and around zero if the jth variable is unimportant in predicting the response. For each original variable $\mathbf{x}_j$, we compute W_j statistic as $|\widehat{\mathrm{MDA}}_j| - |\widehat{\mathrm{MDA}}_{j+p}|$, $j = 1, \cdots, p$.

4.3 Simulation results

For each method, we use 100 simulated data sets to calculate its empirical FDR and power, which are the average FDP and TDP (true discovery proportion as in (5)) over 100 repetitions, respectively. Two different thresholds, knockoff and knockoff+ (T_1 and T_2 in Result 1, respectively), are used in the knockoffs inference implementation. It is worth mentioning that as shown in [15] and summarized in Result 1, knockoff+ controls FDR (3) exactly while knockoff controls only the modified FDR (4).

Tables 1 and 2 summarize the results from Designs 1 and 2, respectively. As shown in Table 1, all approaches can control empirical FDR at the target level ($q = 0.2$) and knockoff+, which is more conservative, reduces power negligibly. It is worth mentioning that even for Design 2, in which the design matrix $\mathbf{X}$ is drawn from fat-tail distribution with serial dependence, we still have FDR under control with decent level of power. This suggests that the no serial correlation assumption in our theoretical analysis could just be technical. Compared to the results by BCKnockoff and HD-BCKnockoff, we see that using the extra information from the factor structure in constructing knockoff variables can help with both FDR and power. Table 2 also shows the effects of model sparsity on the performance of various approaches. It can be seen that when the number of true signals is increased from 50 to 150, the FDR is still under control and the empirical power of IPAD remains steady.

Table 3 is devoted to the case of Design 3, where the rows of matrix $\mathbf{X}$ are generated independently from multivariate normal distribution with AR(1) correlation structure. This is a setting where the factor model structure in $\mathbf{X}$ is misspecified. Since BCknockoff and HD-BCknockoff make no use of the factor structure in generating knockoff variables, in both low- and high-dimensional examples both methods control FDR exactly at the target level. IPAD based methods have empirical FDR slightly over the target level, which may be caused by the misspecification of the factor structure. On the other hand, IPAD based approaches have much higher empirical power than comparison methods.

Table 4 corresponds to Design 4 in which response $\mathbf{y}$ is related to $\mathbf{X}$ nonlinearly. Since BCKnockoff and HD-BCKnockoff are designed for linear models, only the results from IPAD

Table 1: Simulation results for Designs 1 and 2 of Section 4.1 with different values of dimensionality p

	Design 1					Design 2				
	FDR	Power	FDR$_+$	Power$_+$	R^2	FDR	Power	FDR$_+$	Power$_+$	R^2
	$p = 1000$									
IPAD	0.195	0.991	0.180	0.990	0.659	0.199	0.961	0.180	0.960	0.652
BCKnockoff	0.207	0.942	0.192	0.938	0.659	0.172	0.887	0.152	0.885	0.653
	$p = 2000$									
IPAD	0.194	0.979	0.179	0.979	0.649	0.199	0.935	0.183	0.933	0.656
HD-BCKnockoff	0.142	0.706	0.127	0.691	0.649	0.136	0.607	0.113	0.581	0.644
	$p = 3000$									
IPAD	0.191	0.964	0.176	0.963	0.652	0.188	0.913	0.171	0.911	0.658
HD-BCKnockoff	0.172	0.668	0.149	0.658	0.652	0.125	0.559	0.099	0.524	0.651

Note that FDR$_+$ and Power$_+$ are the values of FDR and Power corresponding to the knockoff+ threshold T_2.

Table 2: Simulation results for Designs 1 and 2 of Section 4.1 with different sparsity level s

	Design 1					Design 2				
	FDR	Power	FDR$_+$	Power$_+$	R^2	FDR	Power	FDR$_+$	Power$_+$	R^2
	$s = 50$									
IPAD	0.194	0.979	0.179	0.979	0.649	0.199	0.935	0.183	0.933	0.656
HD-BCKnockoff	0.142	0.706	0.127	0.691	0.649	0.136	0.607	0.113	0.581	0.644
	$s = 100$									
IPAD	0.191	0.978	0.183	0.977	0.783	0.181	0.937	0.174	0.936	0.789
HD-BCKnockoff	0.152	0.703	0.140	0.698	0.787	0.106	0.583	0.097	0.573	0.778
	$s = 150$									
IPAD	0.183	0.973	0.178	0.972	0.842	0.188	0.935	0.182	0.935	0.848
HD-BCKnockoff	0.139	0.660	0.130	0.654	0.858	0.115	0.578	0.106	0.570	0.843

Table 3: Simulation results for Design 3 of Section 4.1

	$\rho = 0$					$\rho = 0.5$				
	FDR	Power	FDR$_+$	Power$_+$	R^2	FDR	Power	FDR$_+$	Power$_+$	R^2
	$p = 1000$									
IPAD	0.204	0.995	0.189	0.995	0.444	0.226	0.984	0.216	0.984	0.446
BCKnockoff	0.188	0.919	0.172	0.917	0.444	0.137	0.827	0.117	0.821	0.445
	$p = 2000$									
IPAD	0.203	0.993	0.189	0.993	0.447	0.220	0.982	0.202	0.980	0.445
HD-BCKnockoff	0.151	0.630	0.126	0.603	0.449	0.115	0.522	0.090	0.467	0.442
	$p = 3000$									
IPAD	0.225	0.988	0.205	0.987	0.445	0.219	0.979	0.206	0.978	0.443
HD-BCKnockoff	0.150	0.589	0.126	0.560	0.446	0.092	0.439	0.064	0.381	0.447

method are reported. It can be seen form Table 4 that IPAD approach can control FDR with reasonably high power even in the nonlinear setting. We also observe that in nonlinear setting, the power of IPAD deteriorates faster as dimensionality p increases compared to the linear setting.

Table 4: Simulation results for Design 4 of Section 4.1

	$\theta = 1$					$\theta = 2$				
	FDR	Power	FDR_+	Power_+	R^2	FDR	Power	FDR_+	Power_+	R^2
	$p = 50$									
$c = 0.1$	0.109	0.839	0.081	0.720	0.707	0.110	0.943	0.061	0.858	0.707
$c = 0.2$	0.137	0.847	0.068	0.726	0.547	0.097	0.920	0.061	0.837	0.547
$c = 0.3$	0.137	0.765	0.091	0.582	0.451	0.123	0.907	0.076	0.774	0.451
	$p = 250$									
$c = 0.1$	0.189	0.740	0.104	0.504	0.702	0.174	0.876	0.139	0.788	0.702
$c = 0.2$	0.218	0.666	0.131	0.522	0.552	0.209	0.831	0.118	0.660	0.552
$c = 0.3$	0.200	0.569	0.101	0.361	0.451	0.224	0.766	0.141	0.599	0.451
	$p = 500$									
$c = 0.1$	0.243	0.661	0.169	0.497	0.702	0.223	0.831	0.173	0.740	0.702
$c = 0.2$	0.204	0.507	0.111	0.266	0.543	0.216	0.749	0.126	0.594	0.543
$c = 0.3$	0.247	0.478	0.128	0.299	0.451	0.241	0.691	0.156	0.550	0.451

5 Empirical analysis

Our simulation results in Section 4 suggest that IPAD is a powerful approach with asymptotic FDR control. We further examine the application of IPAD to the quarterly data on 109 macroeconomic variables from the third quarter of year 1960 (1960Q3) to the fourth quarter of year 2008 (2008Q4) in the United States discussed in [37]. These variables are transformed by taking logarithms and/or differencing following [37]. Our real data analysis consists of two parts. In the first part, we focus on the performance of IPAD method in terms of empirical FDR and power. In the second part, the forecasting performance of IPAD method will be evaluated.

5.1 Simulation study

To evaluate the performance of IPAD approach in terms of empirical FDR and power with real economic data, we set up one additional Monte Carlo simulation study. In this design, we use the transformed macroeconomic variables described above as the design matrix $\mathbf{X}$, but simulate response $\mathbf{y}$ from the model in Design 1 in Section 4.1. We set the number of true signals, the amplitude of signals, and the target FDR level to $s = 10$, $A = 4$, and $q = 0.2$, respectively.

Table 5 shows the results for IPAD and HD-BCKnockoff approaches. As expected, HD-BCKnockoff can control FDR but suffers from lack of power. On the other hand, IPAD has empirical FDR slightly higher than the target level ($q = 0.2$) while its power is reasonably high. These results are consistent with our theory in Section 3 because IPAD only controls

FDR asymptotically. Additional reason for having slightly higher FDR than the target level can be deviation of the design matrix from our factor model assumption. Overall this simulation study indicates that IPAD can control FDR at around the target level with reasonably high power when we use the macroeconomic data set. In the next section, using the same data set we will compare the forecasting performance of IPAD with that of some commonly used forecasting methods in the literature.

Table 5: Real data simulation results with $(n, p) = (195, 109)$

	FDR	Power	FDR$_+$	Power$_+$	R^2
			$c = 0.2$		
IPAD	0.278	0.812	0.223	0.796	0.747
HD-BCKnockoff	0.096	0.009	0.010	0.002	0.758
			$c = 0.3$		
IPAD	0.280	0.757	0.221	0.723	0.665
HD-BCKnockoff	0.149	0.121	0.027	0.036	0.678
			$c = 0.5$		
IPAD	0.286	0.661	0.215	0.571	0.560
HD-BCKnockoff	0.119	0.009	0.008	0.001	0.554

5.2 Forecasting results

In this section, we apply the IPAD approach to the real economic data set for forecasting. One-step ahead prediction is conducted using rolling window of size 120. More specifically, one of the 109 variables is chosen as the response and the remaining 108 variables are treated as predictors. For each quarter between 1990Q3 and 2008Q4, we use the previous 120 periods for model fitting and then one-step ahead prediction is conducted based on the fitted model. We compare the following different methods, where each method is implemented in a same way as IPAD for one-step ahead prediction.

1) Autoregression of order one (AR(1)). Assume that

$$y_t = \alpha_0 + \rho y_{t-1} + \varepsilon_t,$$

where y_t is regressed on y_{t-1}, and α_0 and ρ are the AR(1) coefficients that need to be estimated. With the ordinary least squares estimates $\hat{\alpha}_0$ and $\hat{\rho}$, the one-step ahead prediction based on this model is $\hat{y}_{T+1} = \hat{\alpha}_0 + \hat{\rho} y_T$.

2) Factor augmented AR(1) (FAR). We first extract m factors $\mathbf{f}_1, \cdots, \mathbf{f}_m$ form the 109 transformed macroeconomic variables by principal component analysis (PCA). Denote by $\tilde{\mathbf{f}}_t \in \mathbb{R}^m$ the factor vector at time t extracted from the rows of matrix $[\mathbf{f}_1, \cdots, \mathbf{f}_m] \in \mathbb{R}^{n \times m}$. Then we regress y_t on y_{t-1} and $\tilde{\mathbf{f}}_{t-1}$ and fit the following model

$$y_t = \alpha_0 + \rho y_{t-1} + \boldsymbol{\gamma}' \tilde{\mathbf{f}}_{t-1} + \varepsilon_t$$

with $\boldsymbol{\gamma} \in \mathbb{R}^m$. The number of factors m is determined using the PC_{p1} criterion in [3]. Similar to AR(1) model, one-step ahead forecast of y_t at time T is

$$\hat{y}_{T+1} = \hat{\alpha}_0 + \hat{\rho} y_T + \hat{\boldsymbol{\gamma}}' \tilde{\mathbf{f}}_T.$$

3) Lasso method. The y_t is regressed on y_{t-1}, $\tilde{\mathbf{f}}_{t-1}$, and the 108 transformed macroeconomic variables $\mathbf{z}_{t-1} \in \mathbb{R}^{108}$ at time $t-1$

$$y_t = \alpha_0 + \rho y_{t-1} + \boldsymbol{\gamma}' \tilde{\mathbf{f}}_{t-1} + \boldsymbol{\delta}' \mathbf{z}_{t-1} + \varepsilon_t,$$

where $\tilde{\mathbf{f}}_t$ is the same as in the FAR(1) model, and α_0, ρ, and $\boldsymbol{\delta} \in \mathbb{R}^{108}$ are regression coefficients that need to be estimated. The coefficients are estimated by Lasso method with regularization parameter chosen by the cross-validation. With the estimated Lasso coefficient vector $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$, one-step ahead forecast of y_t at time T is

$$\hat{y}_{T+1} = \hat{\boldsymbol{\beta}}'_{\text{Lasso}} \mathbf{x}_T,$$

where $\mathbf{x}_T$ is the augmented predictor vector at time T.

4) IPAD method. We regress y_t on the augmented vector $(y_{t-1}, \mathbf{z}'_{t-1})'$. The lagged variable y_{t-1} is assumed to be always in the model. To account for this, we implement IPAD in three steps. First, we regress y_t on y_{t-1} and obtain the residuals $e_{y,t}$. Second, we regress each of the 108 variables in $\mathbf{z}_{t-1}$ on y_{t-1} and obtain the residual vector $\mathbf{e}_{z,t-1}$. At last, we fit model (1)–(2) using the IPAD approach by treating $e_{y,t}$ as the response and $\mathbf{e}_{z,t-1}$ as predictors, which returns us a set of selected variables (a subset of the 108 macroeconomic variables). With the set of variables $\widehat{\mathcal{S}}$ selected by IPAD, we fit the following model by the least-squares regression

$$y_t = \alpha_0 + \rho y_{t-1} + \boldsymbol{\delta}' \mathbf{z}_{t-1,\widehat{\mathcal{S}}} + \varepsilon_t, \tag{19}$$

where $\mathbf{z}_{t,\widehat{\mathcal{S}}}$ stands for the subvector of $\mathbf{z}_t$ corresponding to the set of variables $\widehat{\mathcal{S}}$ selected by IPAD at time t. Since $\widehat{\mathcal{S}}$ from IPAD is random due to the randomness in generating knockoff variables, we apply the IPAD procedure 100 times and compute the average of these 100 one-step ahead predictions based on (19) and use the mean value as the final predicted value of y_{T+1}.

Table 6 shows the root mean-squared prediction error (RMSE) of these methods. As can be seen, the RMSE of IPAD is very close to those of comparison methods. To statistically compare the relative prediction accuracy of IPAD versus other approaches, we have used the Diebold–Mariano test [21], where the square of one-step ahead prediction error is used as the loss function. Table 7 reports the test results. The results indicate that one-step ahead prediction accuracy of IPAD is comparable to other approaches.

It is worth mentioning that one main advantage of IPAD is its interpretability and stability. Using IPAD for forecasting, we not only enjoy the same level of accuracy as other methods but also obtain the information on variable importance with stability. Recall that

Table 6: Root mean-squared error of one-period ahead forecast of various macroeconomic variables

	AR	FAR	Lasso	IPAD
RGDP	2.245	1.929	2.070	2.106
CPI-ALL	1.526	1.552	1.579	1.571
Imports	7.549	5.871	6.595	6.993
IP: cons dble	9.683	8.353	8.424	9.175
Emp: TTU	1.112	0.989	1.167	1.100
U: mean duration	0.573	0.487	0.502	0.494
HStarts: South	0.074	0.071	0.076	0.074
NAPM new ordrs	4.800	4.378	4.659	4.673
PCED-NDUR-ENERGY	31.927	32.121	33.546	32.164
Emp. Hours	2.102	1.899	2.080	1.944
FedFunds	0.421	0.396	0.406	0.392
Cons credit	2.573	2.537	2.648	2.580
EX rate: Canada	10.132	10.139	10.122	10.113
DJIA	23.117	23.997	24.585	23.398
Consumer expect	6.496	6.888	6.681	6.661

Table 7: Diebold–Mariano test for comparing prediction accuracy of IPAD against other procedures

	IPAD vs. AR	IPAD vs. FAR	IPAD vs. Lasso
RGDP	-0.780	1.160	0.462
CPI-ALL	0.521	0.394	-0.218
Imports	-0.976	2.631**	1.464
IP: cons dble	-1.026	1.567	2.487*
Emp: TTU	-0.140	1.692	-1.845
U: mean duration	-3.383***	0.672	-0.505
HStarts: South	0.096	0.821	-0.766
NAPM new ordrs	-0.517	1.814	0.076
PCED-NDUR-ENERGY	0.753	0.049	-1.759
Emp. Hours	-1.200	0.297	-2.063*
FedFunds	-0.971	-0.134	-0.625
Cons credit	0.207	0.359	-0.661
EX rate: Canada	-0.466	-0.138	-0.037
DJIA	0.585	-0.959	-1.428
Consumer expect	1.212	-1.038	-0.277

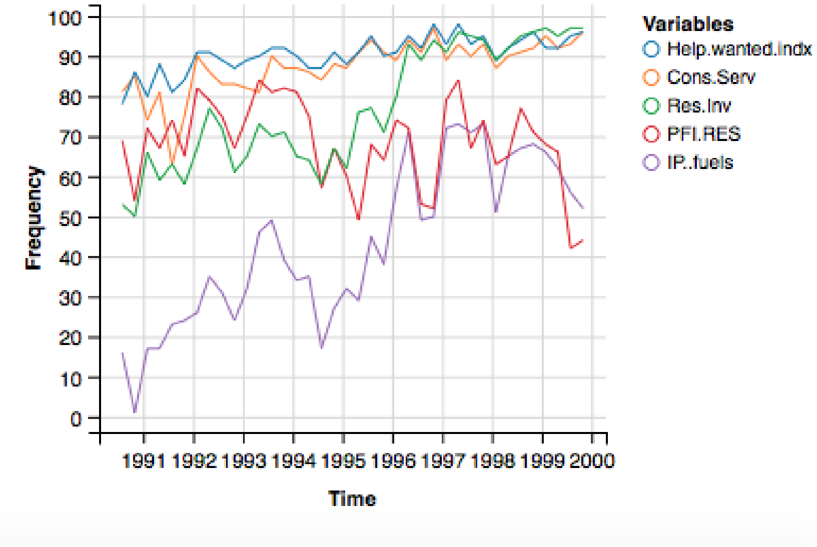
for each one-step ahead prediction, we apply IPAD 100 times and obtain 100 sets of selected variables. Thus we can calculate the selection frequency of each variable in each one-step ahead prediction. Figure 1 depicts the frequencies of top five selected variables in predicting real GDP growth before and after year 2000, where the variable importance is ranked according to the aggregated frequencies over the entire time period before or after 2000. We have experimented with different cutoff years around year 2000, and the top five ranked variables stay the same so only the results corresponding to cutoff year 2000 are reported. Changes in index of help wanted advertising in newspapers, percentages of changes in real personal consumption of services, and percentage of changes in real gross private domestic investment in residential sector were the top three important variables in predicting real GDP growth during the whole period. It is interesting to see that percentage of changes in residential price index was among top five important variables in predicting GDP growth during the 90s, and then starting from year 2000 it was replaced by changes in index of consumer expectations about stability of economy. Moreover, it is also seen that the percentage of changes in industrial production of fuels was of great importance for predicting real GDP growth during some periods but not the others.

As a comparison, it is very difficult to interpret the results of FAR. As for Lasso based method, there is no theoretical guarantee on FDR control and in addition, Lasso usually gives us models with much larger size. For instance, in predicting real GDP growth, IPAD on average selects 5.42 macroeconomic variables while Lasso on average selects 13.32 variables. To summarize, our real data analysis indicates that IPAD is an applicable approach for controlling FDR with competitive prediction power and high interpretability and stability.

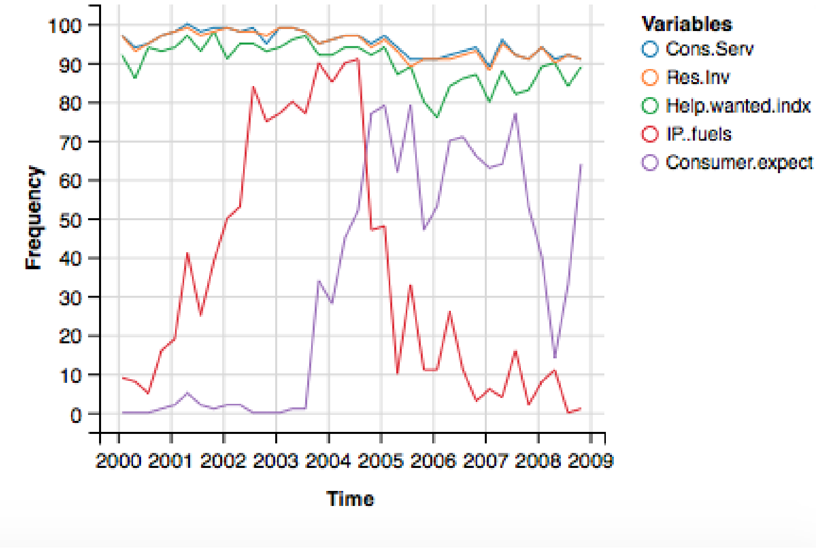
6 Discussions

We have suggested in this paper a new procedure IPAD for feature selection in high-dimensional linear models that achieves asymptotic FDR control while retaining high power. Our model setting involves a latent factor model that is motivated by applications in economics and finance. Our method falls in the general model-X knockoffs framework in [15], but allows the unknown covariate distribution for the knockoff variable construction. With the LCD knockoff statistics, we have shown that the FDR of IPAD can be asymptotically under control while the power can be asymptotically the same as that of Lasso. Our simulation study and empirical analysis also suggest that IPAD has highly competitively performance compared to many widely used forecasting methods such as Lasso and FAR, but with much higher interpretability and stability.

Our work has focused on the scenario of static models. It would be interesting to extend the IPAD procedure to high-dimensional dynamic models with time series data. It is also interesting to consider nonlinear models and more flexible machine learning methods for forecasting as well as more refined factor model structures on the covariates for the knockoffs inference with IPAD, and develop theoretical guarantees for the IPAD framework in these more general model settings. These extensions are beyond the scope of the current paper and are interesting topics for future research.



(a) 1990-1999



(b) 2000-2008

Figure 1: Frequencies of top selected variables in predicting real GDP growth. The set of selected variables are index of help-wanted advertising in newspapers (Help wanted indx), real personal consumption expenditures - services (Cons-Serv), real gross private domestic investment - residential (Res.Inv), residential price index (PFI-RES), industrial production index - fuels (IP:fuels), and University of Michigan index of consumer expectations (Consumer expect).

References

[1] Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*, 1203–1227.

[2] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*,

135–171.

[3] Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

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

[5] Barber, R. F. and E. J. Candès (2016). A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*.

[6] Barber, R. F., E. J. Candès, and R. J. Samworth (2018). Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*.

[7] Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint arXiv:1806.01888*.

[8] Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society Series B 72*, 405–416.

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

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

[11] Bercu, B., B. Delyon, and E. Rio (2015). *Concentration Inequalities for Sums and Martingales (1st ed.).* Springer.

[12] Billingsley, P. (1995). *Probability and Measure (3rd ed.).* Wiley-Interscience.

[13] Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, 13–60.

[14] Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

[15] Candès, E. J., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: 'modelX' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B 80*, 551–577.

[16] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*, C1–C68.

[17] Chernozhukov, V., W. K. Härdle, C. Huang, and W. Wang (2018). Lasso-driven inference in time and space. *arXiv preprint arXiv:1806.05081*.

[18] Chernozhukov, V., W. Newey, and J. Robins (2018). Double/de-biased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*.

[19] Chudik, A., G. Kapetanios, and H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica, to appear*.

[20] De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*, 318–328.

[21] Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*, 134–144.

[22] Durrett, R. (2010). *Probability: Theory and Examples (4th ed.).* Cambridge University Press.

[23] Fan, J. and Y. Fan (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics 36*, 2605–2637.

[24] Fan, J., X. Han, and W. Gu (2012). Estimating false discovery proportion under arbitrary covariance dependence (with discussion). *Journal of American Statistical Association 107*, 1019–1045.

[25] Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B 70*, 849–911.

[26] Fan, Y., E. Demirkaya, G. Li, and J. Lv (2018). RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association, to appear*.

[27] Fan, Y., E. Demirkaya, and J. Lv (2017). Nonuniformity of p-values can occur early in diverging dimensions. *arXiv preprint arXiv:1705.03604*.

[28] Fan, Y. and J. Lv (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association 108*, 1044–1061.

[29] Guo, Z., H. Kang, T. T. Cai, and D. S. Small (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B 80*, 793–815.

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

[31] Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis (2nd ed.).* Cambridge University Press.

[32] Lv, J. (2013). Impacts of high dimensionality in finite samples. *The Annals of Statistics 41*, 2236–2262.

[33] Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science 27*, 538–557.

[34] Rigollet, P. and J.-C. Hütter (2017). *High Dimensional Statistics.* Massachusetts Institute of Technology, MIT Open CourseWare.

[35] Romano, J. P. and M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*, 94–108.

[36] Shah, R. D. and P. Bühlmann (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society Series B 80*, 113–135.

[37] Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*, 481–493.

[38] Stucky, B. and S. van de Geer (2018). Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Transactions on Signal Processing 66*, 2178–2190.

[39] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*, 267–288.

[40] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok (Eds.), *Compressed Sensing: Theory and Practice*, pp. 210–268. Cambridge University Press.

[41] Vizcarra, A. B. and F. G. Viens (2007). Some applications of the Malliavin calculus to sub-Gaussian and non-sub-Gaussian random fields. In R. C. Dalang, M. Dozzi, and F. Russo (Eds.), *Seminar on Stochastic Analysis, Random Fields and Applications V*, pp. 363–395. Springer Science & Business Media.

[42] Wooldridge, J. M. and Y. Zhu (2018). Inference in approximately sparse correlated random effects probit models. *Journal of Bussiness & Economic Statistics, to appear*.

[43] Zhang, X. and G. Cheng (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association 112*, 757–768.

Appendix

This appendix contains all the proofs and technical details for the theoretical results of the paper. In particular, Section A details the proofs of Lemmas 1–2 and Theorems 1–2, Section B presents some key lemmas and their proofs, and Section C provides some additional technical lemmas and their proofs.

To ease the technical presentation, let us introduce some notation. We denote by $\lesssim$ the inequality up to some positive constant factor. Restricting the columns of $\mathbf{X}$ and $\widetilde{\mathbf{X}}(\hat{\boldsymbol{\theta}})$ to the variables in index set $\mathcal{A}$ such that $|\mathcal{A}| \leq k$, we obtain the $n \times k$ submatrices $\mathbf{X}_{\mathcal{A}}$ and $\widetilde{\mathbf{X}}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})$, respectively. Moreover, we define $\mathbf{T}_{\mathcal{A}}(\hat{\boldsymbol{\theta}}) := \mathrm{vec}(\mathrm{vech}\, \mathbf{U}_{\mathcal{A}}(\hat{\boldsymbol{\theta}}), \mathbf{v}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})) \in \mathbb{R}^{k(2k+3)}$ with $\mathbf{U}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})$ the principle submatrix of $\mathbf{U}(\hat{\boldsymbol{\theta}})$ formed by columns and rows in set $\mathcal{A}$, and $\mathbf{v}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})$ the subvector of $\mathbf{v}(\hat{\boldsymbol{\theta}})$ formed by components in set $\mathcal{A}$. Then it is easy to see that $\mathbf{U}_{\mathcal{A}}(\hat{\boldsymbol{\theta}}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})]'[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})]$ and $\mathbf{v}_{\mathcal{A}}(\hat{\boldsymbol{\theta}}) = n^{-1}[\mathbf{X}_{\mathcal{A}}, \widetilde{\mathbf{X}}_{\mathcal{A}}(\hat{\boldsymbol{\theta}})]'\mathbf{y}$. For the oracle factor loading matrix $\boldsymbol{\Lambda}^0$, with a slight abuse of notation we use $\boldsymbol{\Lambda}^0_{\mathcal{A}}$ to denote the *row* restricted to the variables in $\mathcal{A}$ for notational convenience. Recall that $\nu > 0$ is a fixed positive number, $c_{np} = (p^{-1}\log n)^{1/2} + (n^{-1}\log p)^{1/2}$, and $\tilde{c}_{np} = p^{-1/2}\log n + n^{-1/2}\log p$. We define $\pi_{np} = n^{-\nu} + p^{-\nu}$. Since λ is fixed at $C_0 n^{-1/2}\log p$, in all the proofs we will drop the dependence of various quantities on λ whenever there is no confusion.

A Proofs of main results

A.1 Proof of Lemma 1

For λ fixed at $C_0 n^{-1/2}\log p$ and each given $\boldsymbol{\theta}$, $W_j(\boldsymbol{\theta}) = w_j([\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})], \mathbf{y})$ depends only on $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$ by the LCD construction. Moreover, the Lasso solution $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$ satisfies the Karush–Kuhn–Tucker (KKT) conditions:

$$\mathbf{v}(\boldsymbol{\theta}) - \mathbf{U}(\boldsymbol{\theta})\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) = n^{-1}\lambda\mathbf{z}, \tag{A.1}$$

$$\text{where } \mathbf{z} = (z_1, \cdots, z_{2p})^T \text{ with } z_j \in \begin{cases} \{\mathrm{sgn}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \hat{\beta}_j = 0, \end{cases} \quad \text{for} \quad j = 1, \ldots, 2p. \tag{A.2}$$

This means that $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$ depends on the data $([\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})], \mathbf{y}))$ only through $\mathbf{U}(\boldsymbol{\theta})$ and $\mathbf{v}(\boldsymbol{\theta})$. Thus using notation $\mathbf{T}(\boldsymbol{\theta}) = \mathrm{vec}(\mathrm{vech}\, \mathbf{U}(\boldsymbol{\theta}), \mathbf{v}(\boldsymbol{\theta}))$ with the fact that $\mathbf{U}(\boldsymbol{\theta})$ is symmetric, we can reparametrize $w_j([\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})], \mathbf{y})$ as $w_j(\mathbf{T}(\boldsymbol{\theta}))$ with a slight abuse of notation. Furthermore, note that the thresholds T_1 and T_2 are both completely determined by $w_j(\mathbf{T}(\boldsymbol{\theta}))$. Consequently, by the construction of $\widehat{\mathcal{S}}$ we can see that $\widehat{\mathcal{S}}$ depends only on $\mathbf{T}(\boldsymbol{\theta})$, which completes the proof of Lemma 1.

A.2 Proof of Lemma 2

We continue to use the same λ and $\boldsymbol{\theta}$ as in Lemma 1 and its proof. Recall that $S_{\mathcal{A}}(\mathbf{t}_{\mathcal{A}})$ represents the outcome of first restricting ourselves to the smaller set of variables $\mathcal{A}$ and then applying IPAD to $\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) = \mathbf{t}_{\mathcal{A}}$ to further select variables from $\mathcal{A}$. Also recall that $\mathcal{A}^*(\boldsymbol{\theta})$ is the support of knockoff statistics $W_j(\boldsymbol{\theta})$. Thus the knockoff threshold T_1 or T_2 depends only on $W_j(\boldsymbol{\theta})$ with $j \in \mathcal{A}^*(\boldsymbol{\theta})$.

On the other hand, when we restrict ourselves to $\mathcal{A} \supset \mathcal{A}^*(\boldsymbol{\theta})$ we solve the following KKT conditions with respect to $\tilde{\boldsymbol{\beta}} := (\tilde{\beta}_1, \cdots, \tilde{\beta}_{2|\mathcal{A}|})^T \in \mathbb{R}^{2|\mathcal{A}|}$ to get the Lasso solution:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta})'\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}))^{-1}(\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - n^{-1}\lambda\tilde{\mathbf{z}}), \tag{A.3}$$

$$\text{where } \tilde{\mathbf{z}} = (\tilde{z}_1, \cdots, \tilde{z}_{2|\mathcal{A}|})^T \text{ with } \tilde{z}_j \in \begin{cases} \{\operatorname{sgn}(\tilde{\beta}_j)\} & \text{if } \tilde{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \tilde{\beta}_j = 0, \end{cases} \quad \text{for} \quad j = 1, \ldots, 2|\mathcal{A}|. \tag{A.4}$$

Since λ is always fixed at the same value $C_0 n^{-1/2} \log p$, it is seen that the solution to the above KKT conditions is identical to $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}_{\mathcal{A}\mathcal{A}}(\boldsymbol{\theta})$, where the latter denotes the subvector of $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$ formed by stacking $\hat{\beta}^{\mathsf{aug}}_{j_1}(\boldsymbol{\theta})$, $j_1 \in \mathcal{A}$ and $\hat{\beta}^{\mathsf{aug}}_{p+j_2}(\boldsymbol{\theta})$, $j_2 \in \mathcal{A}$ all together. Therefore, the Lasso solution to (A.3)–(A.4) and the Lasso solution to (A.1)–(A.2) have the identical support (when viewed in the original $2p$-dimensional space) and in addition, identical values on the support. This guarantees that $S_{\{1,\ldots,p\}}(\mathbf{T}(\boldsymbol{\theta}))$ and $S_{\mathcal{A}}(\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}))$ are identical and thus concludes the proof of Lemma 2.

A.3 Proof of Theorem 1

Recall that for a given $\boldsymbol{\theta}$, $\mathcal{A}^*(\boldsymbol{\theta})$ is the support of knockoff statistics $(W_1(\boldsymbol{\theta}), \cdots, W_p(\boldsymbol{\theta}))'$. Define set

$$\widehat{\mathcal{A}}(\hat{\boldsymbol{\theta}}) := \mathcal{A}^*(\hat{\boldsymbol{\theta}}) \cup \mathcal{A}^*(\boldsymbol{\theta}^0).$$

It follows from (14) that the cardinality of $\widehat{\mathcal{A}}(\hat{\boldsymbol{\theta}})$ is bounded by k. Hereafter we write $\widehat{\mathcal{A}}(\hat{\boldsymbol{\theta}})$ as $\widehat{\mathcal{A}}$ for notational simplicity.

By Lemmas 1–2 and the definition of the FDP, we know that $S_{\{1,\ldots,p\}}(\mathbf{T}(\hat{\boldsymbol{\theta}})) = S_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}))$ and thus the resulting FDR's are the same. Therefore, we can restrict ourselves to the smaller model $\widehat{\mathcal{A}}$ when studying the FDR of IPAD. The same arguments as above also hold for the oracle knockoffs; that is, the FDR of IPAD applied to $\mathbf{T}(\boldsymbol{\theta}^0)$ is the same as that applied to $\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0)$. Note that all the FDR's we discuss here are with respect to the full model $\{1, \cdots, p\}$. For this reason, in what follows we will abuse the notation and use $\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}))$ and $\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))$ to denote the FDR of IPAD based on $\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta})$ and $\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0)$, respectively. We want to emphasize that although we put a subscript $\widehat{\mathcal{A}}$ in FDR's, their values are still deterministic as argued above. Summarizing the facts, we obtain

$$\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}})) = \mathrm{FDR}_{\{1,\cdots,p\}}(\mathbf{T}(\hat{\boldsymbol{\theta}})),$$
$$\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0)) = \mathrm{FDR}_{\{1,\cdots,p\}}(\mathbf{T}(\boldsymbol{\theta}^0)).$$

Meanwhile, by construction $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ satisfies the two properties in Definition 1 and is a valid model-X knockoffs matrix. Therefore, for any value of the regularization parameter, the LCD statistics $W_j(\boldsymbol{\theta}^0)$ based on $([\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)], \mathbf{y})$ together with Result 1 ensure the exact

FDR control at some target level $q \in (0,1)$. Summarizing this, we obtain that the FDR of IPAD applied to $\mathbf{T}(\boldsymbol{\theta}^0)$ is controlled at target level q.

Combining the arguments in the previous two paragraphs, we deduce

$$\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0)) = \mathrm{FDR}_{\{1,\cdots,p\}}(\mathbf{T}(\boldsymbol{\theta}^0)) \leq q.$$

Thus the desired results follow automatically if we can prove that $\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}))$ is asymptotically close to $\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))$. We next proceed to prove it.

Recall the definitions of $\mathbb{I}$ and $\mathbb{I}_{\mathcal{A}}$ as in (15). Define the event

$$\mathcal{E}_{np} = \left\{\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}}) \in \mathbb{I}_{\widehat{\mathcal{A}}}\right\} \cap \left\{\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0) \in \mathbb{I}_{\widehat{\mathcal{A}}}\right\}.$$

Lemma 3 in Section B.1 establishes $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{np}$ with probability at least $1 - O(\pi_{np})$ and $\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_{np}$. Hence, Lemma 4 in Section B.2 guarantees that

$$\mathbb{P}\left(\mathcal{E}_{np}^c\right) \leq 2\,\mathbb{P}\left(\sup_{|\mathcal{A}|\leq k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} > a_{np}\right) = O(\pi_{np}), \tag{A.5}$$

where $a_{np} = C_1(k^{1/2} + s^{3/2})\tilde{c}_{np}$ for some constant $C_1 > 0$.

For a given deterministic set $\mathcal{A} \subset \{1,\cdots,p\}$, let $\mathrm{FDP}_{\mathcal{A}}(\cdot)$ be the FDP function corresponding to $\mathrm{FDR}_{\mathcal{A}}(\cdot)$. By the definition of FDP function, we have for any $\mathbf{t}_1, \mathbf{t}_2 \in \mathbb{R}^{|\mathcal{A}|(2|\mathcal{A}|+3)}$,

$$\begin{aligned}\mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_2) - \mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_1) &= \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)|}{|S_{\mathcal{A}}(\mathbf{t}_2)|} - \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)|}\\ &= \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| \cdot (|S_{\mathcal{A}}(\mathbf{t}_1)| - |S_{\mathcal{A}}(\mathbf{t}_2)|)}{|S_{\mathcal{A}}(\mathbf{t}_1)| \cdot |S_{\mathcal{A}}(\mathbf{t}_2)|} + \frac{|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| - |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)|}{|S_{\mathcal{A}}(\mathbf{t}_1)|}.\end{aligned}$$

Further, note that

$$\begin{aligned}&|\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)|/|S_{\mathcal{A}}(\mathbf{t}_2)| \leq 1, \qquad ||S_{\mathcal{A}}(\mathbf{t}_2)| - |S_{\mathcal{A}}(\mathbf{t}_1)|| \leq |S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)|,\\ &\big||\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_2)| - |\mathcal{S}^1 \cap S_{\mathcal{A}}(\mathbf{t}_1)|\big| \leq \big|\{S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)\} \cap \mathcal{S}^1\big|.\end{aligned}$$

Combining the results above yields

$$\begin{aligned}&|\,\mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_2)|\\ &\quad \leq \frac{\big||S_{\mathcal{A}}(\mathbf{t}_1)| - |S_{\mathcal{A}}(\mathbf{t}_2)|\big|}{|S_{\mathcal{A}}(\mathbf{t}_1)|} + \frac{\big|\{S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)\} \cap \mathcal{S}^1\big|}{|S_{\mathcal{A}}(\mathbf{t}_1)|} \leq 2\frac{\big|S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)\big|}{|S_{\mathcal{A}}(\mathbf{t}_1)|}.\end{aligned}$$

Similarly we have

$$|\,\mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_2)| \leq 2\frac{\big|S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)\big|}{|S_{\mathcal{A}}(\mathbf{t}_2)|}.$$

Thus it holds that

$$\begin{aligned}\sup_{|\mathcal{A}|\leq k}\ \sup_{\mathbf{t}_1,\mathbf{t}_2\in\mathbb{I}_{\mathcal{A}}} |\,\mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_2)| &\leq \sup_{|\mathcal{A}|\leq k}\ \sup_{\mathbf{t}_1,\mathbf{t}_2\in\mathbb{I}_{\mathcal{A}}} \frac{\big|S_{\mathcal{A}}(\mathbf{t}_2) \triangle S_{\mathcal{A}}(\mathbf{t}_1)\big|}{|S_{\mathcal{A}}(\mathbf{t}_1)| \wedge |S_{\mathcal{A}}(\mathbf{t}_2)|}\\ &= O(\rho_{np}),\end{aligned} \tag{A.6}$$

where the last two steps are due to Condition 6. Therefore, (A.5) and (A.6) together with the fact that $\mathrm{FDP}(\cdot) \in [0,1]$ entail that

$$\begin{aligned}
&\left|\mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}})) - \mathrm{FDR}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))\right| = \left|\mathbb{E}\,\mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}})) - \mathbb{E}\,\mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))\right| \\
&\quad \le \mathbb{E}\left|\mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}})) - \mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))\right| \\
&\quad \le \mathbb{E}\left[\left|\mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\hat{\boldsymbol{\theta}})) - \mathrm{FDP}_{\widehat{\mathcal{A}}}(\mathbf{T}_{\widehat{\mathcal{A}}}(\boldsymbol{\theta}^0))\right| \mid \mathcal{E}_{np}\right]\mathbb{P}\left(\mathcal{E}_{np}\right) + 2\,\mathbb{P}\left(\mathcal{E}_{np}^c\right) \\
&\quad \le \sup_{|\mathcal{A}|\le k}\ \sup_{\mathbf{t}_1,\mathbf{t}_2\in\mathbb{I}_{\mathcal{A}}} |\mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_1) - \mathrm{FDP}_{\mathcal{A}}(\mathbf{t}_2)| + O(\pi_{np}) \\
&\quad = O(\rho_{np}) + O(\pi_{np}).
\end{aligned}$$

This completes the proof of Theorem 1.

A.4 Proof of Theorem 2

By the definition of the LCD statistics, we construct the augmented Lasso estimator for each $\boldsymbol{\theta} \in \Theta_{np}$, which is defined as

$$\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{b}\in\mathbb{R}^{2p}} \left\|\mathbf{y} - [\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]\mathbf{b}\right\|_2^2 + \lambda\|\mathbf{b}\|_1. \tag{A.7}$$

The Lasso estimator of regressing $\mathbf{y}$ on only $\mathbf{X}$ is also given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}\in\mathbb{R}^{p}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1, \tag{A.8}$$

where $\lambda = O(n^{-1/2}\log p)$. According to the true model $\mathcal{S}^0$, the underlying true parameter vector corresponding to $\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})$ should be given by $\boldsymbol{\beta}^{\mathsf{aug}} := (\boldsymbol{\beta}', \mathbf{0}')' \in \mathbb{R}^{2p}$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{\mathcal{S}^0}, \mathbf{0}')' \in \mathbb{R}^p$ and $|\mathcal{S}^0| = s$ for any $\boldsymbol{\theta} \in \Theta_{np}$. By Lemma 5 in Section B.3, with probability at least $1 - O(\pi_{np})$ the Lasso estimators satisfy

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{\mathsf{aug}}\right\|_1 &= O(s\lambda), \\
\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1 &= O(s\lambda),
\end{aligned}$$

where $\lambda = O(n^{-1/2}\log p)$.

We now prove that under Condition 7, the power of the augmented Lasso (A.7) is bounded from below by $\gamma \in [0,1]$; that is,

$$\mathbb{E}\left|\widehat{\mathcal{S}}_{\mathsf{auglasso}} \cap \mathcal{S}^0\right|/s \ge \gamma, \tag{A.9}$$

where $\widehat{\mathcal{S}}_{\mathsf{auglasso}} = \{j : \hat{\beta}_j^{\mathsf{aug}}(\boldsymbol{\theta}) \neq 0\}$. To this end, we first show that with asymptotic probability one,

$$|\widehat{\mathcal{S}}^c_{\mathsf{auglasso}} \cap \mathcal{S}^0|/s \le 1 - \gamma. \tag{A.10}$$

The key is to use proof by contradiction. Suppose $|\widehat{\mathcal{S}}^c_{\mathsf{auglasso}} \cap \mathcal{S}^0|/s > 1 - \gamma$. Then we can see that

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{\mathsf{aug}}\right\|_1 &\geq \sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\hat{\boldsymbol{\beta}}^{\mathsf{aug}}_{\widehat{\mathcal{S}}^c_{\mathsf{auglasso}}}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{\mathsf{aug}}_{\widehat{\mathcal{S}}^c_{\mathsf{auglasso}}}\right\|_1 \\ &= \left\|\boldsymbol{\beta}^{\mathsf{aug}}_{\widehat{\mathcal{S}}^c_{\mathsf{auglasso}}}\right\|_1 \geq \left\|\boldsymbol{\beta}^{\mathsf{aug}}_{\widehat{\mathcal{S}}^c_{\mathsf{auglasso}}\cap\mathcal{S}^0}\right\|_1 > b_{np} s n^{-1/2}\log p,\end{aligned}$$

where the last step is by Condition 7. However, by Lemma 5 with probability at least $1-O(\pi_{np})$, the left hand side above is bounded from above by $O(s\lambda)$ with $\lambda = O(n^{-1/2}\log p)$. These two results contradict with each other since $b_{np} \to \infty$. Hence (A.10) is proved. Therefore, the result in (A.9) follows immediately since $|\widehat{\mathcal{S}}_{\mathsf{auglasso}} \cap \mathcal{S}^0| = s - |\widehat{\mathcal{S}}^c_{\mathsf{auglasso}} \cap \mathcal{S}^0|$ and

$$\begin{aligned}&\mathbb{E}\left|\widehat{\mathcal{S}}_{\mathsf{auglasso}} \cap \mathcal{S}^0\right|/s \geq \gamma\,\mathbb{P}\left(|\widehat{\mathcal{S}}_{\mathsf{auglasso}} \cap \mathcal{S}^0|/s > \gamma\right) \\ &= \gamma\,\mathbb{P}\left(|\widehat{\mathcal{S}}^c_{\mathsf{auglasso}} \cap \mathcal{S}^0|/s \leq 1-\gamma\right) = \gamma(1 - O(\pi_{np})).\end{aligned}$$

Using the same argument, we can show that the power of the Lasso (A.8) is also bounded from below by $\gamma(1 - O(\pi_{np}))$ under Condition 7. That is, we have

$$\mathbb{E}\left|\widehat{\mathcal{S}}_{\mathsf{lasso}} \cap \mathcal{S}^0\right|/s \geq \gamma(1 - O(\pi_{np})),$$

where $\widehat{\mathcal{S}}_{\mathsf{lasso}} = \{j : \hat{\beta}_j \neq 0\}$.

Next we show that our knockoffs procedure has at least the same power as the augmented Lasso and hence the Lasso itself. Namely, we prove

$$\mathbb{E}\left|\widehat{\mathcal{S}} \cap \mathcal{S}^0\right|/s \geq \gamma \tag{A.11}$$

with threshold T_2. Note that the same argument is still valid for T_1. Let $|W_{(1)}| \geq \cdots \geq |W_{(p)}|$ and define j^* as $|W_{(j^*)}| = T_2$. Then by the definition of T_2, it holds that $-T_2 < W_{j^*+1} \leq 0$. Here we have assumed that there are no ties on the magnitudes of W_j's which should be a reasonable assumption considering the continuity of the Lasso solution. As in the proof of Theorem 3 in [26], it is sufficient to consider the following two cases.

Case 1. Consider the case of $-T_2 < W_{(j^*+1)} < 0$. In this case, from the definition of threshold T_2 we have

$$\frac{2 + |\{j : W_{(j)} \leq -T_2\}|}{|\{j : W_{(j)} \geq T_2\}|} > q.$$

Using the same argument as in Lemma 6 of [26] together with Lemma 5, we can prove from Condition 8 that $|\widehat{\mathcal{S}}| \geq C_2 s$ with probability at least $1 - O(\pi_{np})$. This leads to $|\{j : W_{(j)} \leq -T_2\}| > C_2 qs - 2$ with the same probability. Now from the same argument as in A.5 of [26], we can obtain $T_2 = O(\lambda)$. On the other hand, Lemma 5 and some algebra establish that

$$\begin{aligned}
O(s\lambda) = \|\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\beta}^{\mathsf{aug}}\|_1 &= \sum_{j=1}^{p} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| + \sum_{j=1}^{p} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| \\
&= \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| \\
&\quad + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| + \sum_{j=1}^{p} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})|. \qquad \text{(A.12)}
\end{aligned}$$

We then consider the lower bound of the last term in (A.12). For any $j \in \widehat{S}^c$, it holds that $|\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| > |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| - T_2$. Hence we obtain

$$\begin{aligned}
\sum_{j=1}^{p} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| \\
&\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|. \qquad \text{(A.13)}
\end{aligned}$$

Plugging (A.13) into (A.12) and applying the triangle inequality yield

$$\begin{aligned}
O(s\lambda) &\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| \\
&\quad + \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| \\
&\geq \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}}) - \beta_j| + \sum_{j \in \mathcal{S}^1} |\hat{\beta}_j^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| \\
&\quad + \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\beta_j| + \sum_{j \in \widehat{\mathcal{S}} \cap \mathcal{S}^0} |\hat{\beta}_{j+p}^{\mathsf{aug}}(\hat{\boldsymbol{\theta}})| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| \\
&\geq \sum_{j \in \widehat{\mathcal{S}}^c \cap \mathcal{S}^0} |\beta_j| - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| = \|\boldsymbol{\beta}_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 - T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|.
\end{aligned}$$

Since $T_2 |\widehat{\mathcal{S}}^c \cap \mathcal{S}^0| = O(s\lambda)$ for $\lambda = O(n^{-1/2} \log p)$ due to the discussion above, we consequently obtain

$$\|\boldsymbol{\beta}_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 = O(s n^{-1/2} \log p). \qquad \text{(A.14)}$$

Suppose $|\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|/s > 1 - \gamma$. Then Condition 7 gives $\|\boldsymbol{\beta}_{\widehat{\mathcal{S}}^c \cap \mathcal{S}^0}\|_1 > b_{np} s n^{-1/2} \log p$ for some positive diverging sequence b_{np}; this contradicts with (A.14). Thus we obtain $|\widehat{\mathcal{S}}^c \cap \mathcal{S}^0|/s \leq 1 - \gamma$ with asymptotic probability one, which leads to (A.11) by taking expectation.

Case 2. Consider the case of $W_{(j^*+1)} = 0$. In this case, by the definition of threshold T_2

$$\frac{1 + |\{j : W_{(j)} < 0\}|}{|\{j : W_{(j)} > 0\}|} \leq q. \qquad \text{(A.15)}$$

If $|\{j : W_{(j)} < 0\}| > C_3 s$ for some constant $C_3 > 0$, then from the same argument as in A.5 of [26], we can obtain $T_2 = O(\lambda)$, and the rest of the proof is the same as in Case 1. On the other hand, if $|\{j : W_{(j)} < 0\}| \leq o(s)$ we have

$$\begin{aligned}|\{j : W_{(j)} \neq 0\} \cap \mathcal{S}^0| &= |\{j : W_{(j)} > 0\} \cap \mathcal{S}^0| + |\{j : W_{(j)} < 0\} \cap \mathcal{S}^0| \\ &\leq |\widehat{\mathcal{S}} \cap \mathcal{S}^0| + o(s).\end{aligned}$$

Now note that $|\{j : W_{(j)} \neq 0\}| \geq |\{j : |\hat{\beta}_j^{\mathsf{aug}}| \neq 0,\ j = 1, \cdots, p\}|$. Then we can see that with asymptotic probability one,

$$\begin{aligned}|\{j : W_{(j)} \neq 0\} \cap \mathcal{S}^0| &\geq |\{j : \hat{\beta}_j^{\mathsf{aug}} \neq 0,\ j = 1, \cdots, p\} \cap \mathcal{S}^0| \\ &= |\widehat{\mathcal{S}}_{\mathsf{auglasso}} \cap \mathcal{S}^0|. \\ &\geq \gamma s(1 - o(1)).\end{aligned}$$

Consequently, we obtain $|\widehat{\mathcal{S}} \cap \mathcal{S}^0|/s \geq \gamma(1 - o(1))$, which leads to (A.11) by taking expectation. Combining these two cases concludes the proof of Theorem 2.

B Some key lemmas and their proofs

B.1 Lemma 3 and its proof

Lemma 3 *Assume that Conditions 2–5 hold. Then with probability at least $1 - O(\pi_{np})$, the estimator $\hat{\boldsymbol{\theta}} = (\mathrm{vec}(\widehat{\mathbf{C}})', \hat{\boldsymbol{\eta}}')'$ lies in the shrinking set given by*

$$\Theta_{np} = \left\{\boldsymbol{\theta} = (\mathrm{vec}(\mathbf{C})', \boldsymbol{\eta}')' : \left\|\mathbf{C} - \mathbf{C}^0\right\|_{\max} + \left\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\right\|_{\max} \leq O(c_{np})\right\},$$

where $c_{np} = (n^{-1} \log p)^{1/2} + (p^{-1} \log n)^{1/2}$ and $\pi_{np} = p^{-\nu} + n^{-\nu}$.

Proof. We divide the proof into two parts. We prove the bound for $\|\widehat{\mathbf{C}} - \mathbf{C}^0\|_{\max}$ in Part 1 and then for $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\|_{\max}$ in Part 2.

Part 1. Note that $\|\widehat{\mathbf{C}} - \mathbf{C}^0\|_{\max} = \max_{i,j} |\hat{c}_{ij} - c_{ij}^0|$, where the maximum is taken over $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. We write $\mathbf{f}_i^* = \mathbf{H}'\mathbf{f}_i^0$ and $\boldsymbol{\lambda}_j^* = \mathbf{H}^{-1}\boldsymbol{\lambda}_j^0$ with rotation matrix $\mathbf{H}$ defined in Lemma 6 in Section C.1. From the definition of c_{ij}, it holds that

$$\hat{c}_{ij} - c_{ij}^0 = (\hat{\mathbf{f}}_i - \mathbf{f}_i^*)'\boldsymbol{\lambda}_j^* + \hat{\mathbf{f}}_i'(\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*).$$

From Lemma 6, we can assume $\|\mathbf{H}\|_2 + \|\mathbf{H}^{-1}\|_2 + \|\mathbf{V}\|_2 + \|\mathbf{V}^{-1}\|_2 \lesssim 1$, which occurs with probability at least $1 - O(p^{-\nu})$. We also have $\max_{i \in \{1, \ldots, n\}} \|\hat{\mathbf{f}}_i\|_2^2 \lesssim 1$ a.s. by the assumed restriction $\hat{\mathbf{F}}'\hat{\mathbf{F}}/n = \mathbf{I}_r$ as mentioned on p.213 of [3]. Hence, the triangle and Cauchy–Schwarz inequalities with Conditions 2 and 3 give

$$\begin{aligned}\max_{i,j} |\hat{c}_{ij} - c_{ij}| &\leq \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 \max_j \|\boldsymbol{\lambda}_j^*\|_2 + \max_i \|\hat{\mathbf{f}}_i\|_2 \max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2 \\ &\lesssim \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 + \max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2. \qquad \text{(B.1)}\end{aligned}$$

Then it is sufficient to derive upper bounds for $\max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2$ and $\max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2$ that hold with high probability. Using the decomposition of A.1 in [2] along with taking maximum over $i, \ell \in \{1, \dots, n\}$, we can deduce

$$\begin{aligned}
&\max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 \\
&\quad \le \|\mathbf{V}^{-1}\|_2 \max_i \Bigg((\sigma_e^2/n)\|\hat{\mathbf{f}}_i\|_2 + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell\|_2 \left| p^{-1}\sum_{j=1}^p (e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}])\right| \\
&\qquad + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell \mathbf{f}_\ell^{0\prime}\|_2 \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{ij}\right\|_2 + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell \mathbf{f}_i^{0\prime}\|_2 \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{\ell j}\right\|_2 \Bigg) \\
&\quad \lesssim O(n^{-1}) + \max_{i,\ell}\left| p^{-1}\sum_{j=1}^p (e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}])\right| + \max_i \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{ij}\right\|_2 \\
&\quad \lesssim O(n^{-1}) + R_1 + R_2, \qquad \text{(B.2)}
\end{aligned}$$

where we have used the boundedness of $\|\hat{\mathbf{f}}_\ell\|_2$ discussed above and $\|\mathbf{f}_\ell^0\|_2 \le r^{1/2}\|\mathbf{f}_\ell^0\|_{\max} \lesssim 1$ in Condition 2 for the second inequality, and defined $R_1 = \max_{i,\ell}\left| p^{-1}\sum_{j=1}^p (e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}])\right|$ and $R_2 = \max_{i,k}\left| p^{-1}\sum_{j=1}^p \lambda_{jk}^0 e_{ij}\right|$. Similarly, the expression on p.165 of [2] with taking maximum over $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$ leads to

$$\begin{aligned}
&\max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2 \\
&\quad \le \|\mathbf{H}\|_2 \max_j \left\| n^{-1}\sum_{i=1}^n \mathbf{f}_i^0 e_{ij}\right\|_2 + \left\| n^{-1}\sum_{i=1}^n \hat{\mathbf{f}}_i(\hat{\mathbf{f}}_i - \mathbf{f}_i^*)'\right\|_2 \left\|\mathbf{H}^{-1}\right\|_2 \max_j \left\|\boldsymbol{\lambda}_j^0\right\|_2 \\
&\qquad + \max_j \left\| n^{-1}\sum_{i=1}^n (\hat{\mathbf{f}}_i - \mathbf{f}_i^*)e_{ij}\right\|_2 \\
&\quad \lesssim \max_j \left\| n^{-1}\sum_{i=1}^n \mathbf{f}_i^0 e_{ij}\right\|_2 + \max_i \left\|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\right\|_2 + \max_i \left\|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\right\|_2 \max_j \left(n^{-1}\sum_{i=1}^n e_{ij}^2\right)^{1/2} \\
&\quad = R_3 + \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2(1 + R_4), \qquad \text{(B.3)}
\end{aligned}$$

where $R_3 = \max_{j,k}\left| n^{-1}\sum_{i=1}^n f_{ik}^0 e_{ij}\right|_2$ and $R_4 = \max_j \left(n^{-1}\sum_{i=1}^n e_{ij}^2\right)^{1/2}$, and the Cauchy–Schwarz inequality has been used to obtain the second inequality. To evaluate R_4, we note that

$$R_4^2 \le \max_j \mathbb{E}\, e_{ij}^2 + \max_j \left| n^{-1}\sum_{i=1}^n \left(e_{ij}^2 - \mathbb{E}\, e_{ij}^2\right)\right|.$$

The first term is bounded by $2C_e^2$. For the second term, Lemma 7(a) in Section C.2 with p replaced by n and the union bound give

$$\begin{aligned}
\mathbb{P}\left(\max_j \left| n^{-1}\sum_{i=1}^n \left(e_{ij}^2 - \mathbb{E}\, e_{ij}^2\right)\right| > u\right) &\le p \max_j \mathbb{P}\left(\left| n^{-1}\sum_{i=1}^n \left(e_{ij}^2 - \mathbb{E}\, e_{ij}^2\right)\right| > u\right) \\
&\le 2p\exp(-nu^2/C)
\end{aligned}$$

for all $0 \le u \le c$. Thus putting $u = (C(\nu + 1)n^{-1}\log p)^{1/2}$ and using condition $c_{np} \le c/(r^2M^2C(\nu + 2))^{1/2}$, we obtain $R_4^2 = O(1) + O((n^{-1}\log p)^{1/2}) = O(1)$ with probability at least $1 - O(p^{-\nu})$. This together with the observation from (B.1)–(B.3) yields

$$\begin{aligned}\max_{i,j} |\hat{c}_{ij} - c_{ij}^0| &\lesssim R_3 + \left\{R_1 + R_2 + O(n^{-1})\right\}(1 + R_4) \\ &\lesssim R_1 + R_2 + R_3 + O(n^{-1}).\end{aligned}$$

Hence the convergence rate of $\max_{i,j} |\hat{c}_{ij} - c_{ij}^0|$ is determined by the slowest term out of R_1, R_2, R_3, and $O(n^{-1})$. We evaluate these terms by Lemma 7 in Section C.2 and the union bound with condition $c_{np} \leq c/(r^2 M^2 C(\nu+2))^{1/2}$ as above. First for R_1, Lemma 7(a) by letting $u_1 = (C(\nu+2)p^{-1}\log n)^{1/2}$ results in

$$\mathbb{P}\left(R_1 > u_1\right) \leq 2n^2 \exp\left\{-p(\nu+2)p^{-1}\log n\right\} = O(n^{-\nu}).$$

Next for R_2, Lemma 7(c) with $u_2 = (2(\nu+1)p^{-1}\log n)^{1/2}$ gives

$$\mathbb{P}\left(R_2 > u_2\right) \leq 2rn \exp\left\{-p(\nu+1)p^{-1}\log n\right\} = O(n^{-\nu}).$$

Finally for R_3, Lemma 7(b) with putting $u_3 = (C(\nu+1)n^{-1}\log p)^{1/2}$ leads to

$$\mathbb{P}\left(R_3 > u_3\right) \leq 2rp \exp\left\{-n(\nu+1)n^{-1}\log p\right\} = O(p^{-\nu}).$$

Consequently, we obtain the first result $\|\widehat{\mathbf{C}} - \mathbf{C}^0\|_{\max} = O(c_{np})$, which holds with probability at least $1 - O(\pi_{np})$.

Part 2. Next we derive the convergence rate of $\hat{\boldsymbol{\eta}}$. It is sufficient to prove only the case when $\boldsymbol{\eta}^0$ is a scalar (so that we write $\boldsymbol{\eta}^0 = \eta_1^0$) since dimensionality m is fixed and η_k^0's share the identical property thanks to Condition 4. Recall notation $\mathbb{E}_{np} e^k = (np)^{-1} \sum_{i,j} e_{ij}^k$. Letting $\delta_{ij} = c_{ij}^0 - \hat{c}_{ij}$, we have $\hat{e}_{ij} = x_{ij} - \hat{c}_{ij} = e_{ij} + \delta_{ij}$. For an arbitrary fixed $k \in \{1, \ldots, m\}$, the binomial expansion entails

$$\begin{aligned}\left|\mathbb{E}_{np}\hat{e}^k - \mathbb{E}\, e^k\right| &= \left|\mathbb{E}_{np}(e+\delta)^k - \mathbb{E}\, e^k\right| \\ &= \left|\mathbb{E}_{np}(e^k - \mathbb{E}\, e^k) + \mathbb{E}_{np}\sum_{\ell=0}^{k-1}\binom{k}{\ell} e^\ell \delta^{k-\ell}\right| \\ &\leq \left|\mathbb{E}_{np}(e^k - \mathbb{E}\, e^k)\right| + \sum_{\ell=0}^{k-1}\binom{k}{\ell}\max_{i,j}|\delta_{ij}|^{k-\ell}\mathbb{E}_{np}|e|^\ell \\ &\lesssim \left|\mathbb{E}_{np}(e^k - \mathbb{E}\, e^k)\right| + O\left(\max_{i,j}|\delta_{ij}|\right)\sum_{\ell=0}^{k-1}\mathbb{E}_{np}|e|^\ell. \qquad \text{(B.4)}\end{aligned}$$

For all $k \in \{1, \ldots, m\}$, the strong law of large numbers with Theorem 2.5.7 in [22] entails $|\mathbb{E}_{np}e^k - \mathbb{E}\, e^k| = o((np)^{-1/2}\log(np))$ a.s. under Condition 4. Furthermore, the second term of (B.4) is $O(c_{np})$ with probability at least $1 - O(\pi_{np})$ from Part 1 and the same law of large numbers. Consequently, we obtain

$$\left|\mathbb{E}_{np}\hat{e}^k - \mathbb{E}\, e^k\right| \lesssim c_{np}.$$

Therefore by the construction of $\hat{\eta}_1$ and local Lipschitz continuity of h_1 in Condition 4, we see that

$$\begin{aligned}\left|\hat{\eta}_1 - \eta_1^0\right| &= \left|h_1\left(\mathbb{E}_{np}\hat{e}, \dots, \mathbb{E}_{np}\hat{e}^m\right) - h_1\left(\mathbb{E}\, e, \dots, \mathbb{E}\, e^m\right)\right| \\ &\lesssim \max_{k\in\{1,\dots,m\}} \left|\mathbb{E}_{np}\hat{e}^k - \mathbb{E}\, e^k\right|\end{aligned}$$

with probability at least $1 - O(\pi_{np})$. This completes the proof of Lemma 3.

B.2 Lemma 4 and its proof

Lemma 4 *Assume that Conditions 1–4 hold. Then with probability at least $1 - O(\pi_{np})$, the following statements hold*

$$\begin{aligned}(a) \qquad &\sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} = O\left(k^{1/2}\tilde{c}_{np}\right), \\ (b) \qquad &\sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} = O\left(s^{3/2}\tilde{c}_{np}\right),\end{aligned}$$

where Θ_{np} was defined in Lemma 3 and $\tilde{c}_{np} = n^{-1/2}\log p + p^{-1/2}\log n$. Consequently, we have

$$\sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} = O\left(\left(k^{1/2} + s^{3/2}\right)\tilde{c}_{np}\right).$$

Proof. To complete the proof of (a), we verify the following

$$\begin{aligned}(a\text{–}i) \qquad &\sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} \lesssim k^{1/2}\tilde{c}_{np}, \\ (a\text{–}ii) \qquad &\left\|\mathbf{U}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\right\|_{\max} \lesssim (n^{-1}\log p)^{1/2}.\end{aligned}$$

From $(a\text{–}i)$ and $(a\text{–}ii)$, we can conclude that

$$\begin{aligned}&\sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} \\ &\quad\le \sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} + \sup_{|\mathcal{A}|\le k} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)]\right\|_{\max} \\ &\quad\le \sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} + \left\|\mathbf{U}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\right\|_{\max} \\ &\quad\lesssim k^{1/2}\tilde{c}_{np},\end{aligned}$$

which yields result (a).

We begin with showing $(a\text{–}i)$; this is the uniform extension of Lemma 8(a) in Section C.3 over $|\mathcal{A}| \le k$. In fact, the proof is almost the same, with the only difference that bound (B.15) should be replaced with the bound derived in Lemma 9(c); that is,

$$\max_{|\mathcal{A}|\le k} \left\|n^{-1/2}\mathbf{E}_{\mathcal{A}}\right\|_2 \lesssim 1 \vee \left(kn^{-1}\log p\right)^{1/2}, \tag{B.5}$$

which holds with probability at least $1 - O(p^{-\nu})$. Notice that $\left(kn^{-1}\log p\right)^{1/2} \le \log^{1/2} p$. Therefore, even if we use (B.5) instead of (B.15) in the proof of Lemma 8(a) we can still derive the same convergence rate $k^{1/2}\tilde{c}_{np}$ as in Lemma 8(a), and hence (*a–i*) holds with probability at least $1-O(\pi_{np})$.

For (*a–ii*), we see that

$$\begin{aligned}\left\|\mathbf{U}(\boldsymbol{\theta}^0)-\mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\right\|_{\max} &\le \left\|n^{-1}\mathbf{X}'\mathbf{X}-\mathbb{E}[n^{-1}\mathbf{X}'\mathbf{X}]\right\|_{\max}\\ &+\left\|n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)-\mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)]\right\|_{\max}\\ &+2\left\|n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)-\mathbb{E}[n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)]\right\|_{\max} =: W_1+W_2+2W_3. \end{aligned} \tag{B.6}$$

We derive the bounds for each of these terms. First, W_1 is bounded as

$$\begin{aligned}W_1 &\le \left\|n^{-1}\mathbf{C}^{0\prime}\mathbf{C}^0-\mathbb{E}[n^{-1}\mathbf{C}^{0\prime}\mathbf{C}^0]\right\|_{\max}+\left\|n^{-1}\mathbf{E}'\mathbf{E}-\mathbb{E}\,n^{-1}\mathbf{E}'\mathbf{E}\right\|_{\max}+2\left\|n^{-1}\mathbf{E}'\mathbf{C}^0\right\|_{\max}\\ &=: W_{1,1}+W_{1,2}+W_{1,3}.\end{aligned}$$

Under Condition 3, we deduce

$$\begin{aligned}W_{1,1} &= \max_{j,\ell\in\{1,\ldots,p\}}\left|\sum_{k,m=1}^{r}\lambda^0_{jk}\lambda^0_{\ell m}n^{-1}\sum_{i=1}^{n}\left(f^0_{ik}f^0_{im}-\mathbb{E}\,f^0_{ik}f^0_{im}\right)\right|\\ &\le rM^2\max_{j,\ell\in\{1,\ldots,p\}}\left|n^{-1}\sum_{i=1}^{n}\left(f^0_{ik}f^0_{im}-\mathbb{E}\,f^0_{ik}f^0_{im}\right)\right|.\end{aligned}$$

From Lemma 7(d) with Condition 2 and the union bound, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{j,\ell\in\{1,\ldots,p\}}\left|n^{-1}\sum_{i=1}^{n}\left(f^0_{ik}f^0_{im}-\mathbb{E}\,f^0_{ik}f^0_{im}\right)\right|>u\right)\\ &\le p^2\max_{j,\ell\in\{1,\ldots,p\}}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\left(f^0_{ik}f^0_{im}-\mathbb{E}\,f^0_{ik}f^0_{im}\right)\right|>u\right)\le 2p^2\exp\left(-nu^2/C\right).\end{aligned}$$

Hence, letting $u=(C(\nu+2)n^{-1}\log p)^{1/2}$ above yields the bound $W_{1,1}\lesssim(n^{-1}\log p)^{1/2}$ with probability at least $1-O(p^{-\nu})$. Next for $W_{1,2}$, we can find from Lemma 7(a) with p replaced by n and the union bound that

$$\begin{aligned}&\mathbb{P}\left(\left\|n^{-1}\mathbf{E}'\mathbf{E}-\mathbb{E}\,n^{-1}\mathbf{E}'\mathbf{E}\right\|_{\max}>u\right)\le p^2\max_{j,\ell}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\left(e_{ij}e_{i\ell}-\mathbb{E}\,e_{ij}e_{i\ell}\right)\right|>u\right)\\ &\le 2p^2\exp\left(-nu^2/C\right).\end{aligned}$$

Letting $u=(C(\nu+2)n^{-1}\log p)^{1/2}$ and using $n^{-1}\log p\le c^2/(C(\nu+2))$, we obtain $W_{1,2}\lesssim(n^{-1}\log p)^{1/2}$ with probability at least $1-O(p^{-\nu})$. Next for $W_{1,3}$, the union bound gives

$$\begin{aligned}\mathbb{P}\left(\left\|n^{-1}\mathbf{E}'\mathbf{F}^0\boldsymbol{\Lambda}^{0\prime}\right\|_{\max}>u\right) &= \mathbb{P}\left(\max_{j,\ell\in\{1,\ldots,p\}}\left|n^{-1}\sum_{k=1}^{r}\sum_{i=1}^{n}e_{ij}f^0_{ik}\lambda^0_{\ell k}\right|>u\right)\\ &\le \mathbb{P}\left(r\max_{j,\ell\in\{1,\ldots,p\}}\max_{k\in\{1,\ldots,r\}}\left|n^{-1}\sum_{i=1}^{n}e_{ij}f^0_{ik}\right|\left|\lambda^0_{\ell k}\right|>u\right)\\ &\le rp\max_{k\in\{1,\ldots,r\}}\max_{j\in\{1,\ldots,p\}}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}e_{ij}f^0_{ik}\right|>u/(rM)\right).\end{aligned}$$

Lemma 7(b) states that for all $0 \le u/(rM) \le c/(rM)$ it holds that

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} e_{ij}f_{ik}^{0}\right| > u/(rM)\right) \le 2\exp\left\{-nu^2/(Cr^2M^2)\right\}.$$

Therefore, if we put $u = rM(C(\nu+1)n^{-1}\log p)^{1/2}$ using $n^{-1}\log p \le c^2/(r^2M^2C(\nu+1))$, the upper bound of the probability is further bounded by $2rp^{-\nu}$. Thus we obtain $W_{13} \lesssim (n^{-1}\log p)^{1/2}$ with probability at least $1 - O(p^{-\nu})$. Consequently, the bound of W_1 is

$$W_1 \le W_{1,1} + W_{1,2} + W_{1,3} \lesssim (n^{-1}\log p)^{1/2}$$

with probability at least $1 - O(p^{-\nu})$. Note that we have the same result for W_2 since it has the same distribution as W_1. Finally, W_3 is bounded as

$$\begin{aligned} W_3 \le & \left\|n^{-1}\mathbf{C}^{0\prime}\mathbf{C}^0 - \mathbb{E}[n^{-1}\mathbf{C}^{0\prime}\mathbf{C}^0]\right\|_{\max} + \left\|n^{-1}\mathbf{E}'\mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max} \\ & + \left\|n^{-1}\mathbf{E}'\mathbf{C}^0\right\|_{\max} + \left\|n^{-1}\mathbf{E}_{\boldsymbol{\eta}^0}'\mathbf{C}^0\right\|_{\max} =: W_{1,1} + W_{3,1} + W_{1,3} + W_{3,2}. \end{aligned}$$

The upper bound of $W_{3,1}$ turns out to be $O((n^{-1}\log p)^{1/2})$ that holds with probability at least $1 - O(p^{-\nu})$. We check this claim. Using the union bound and the inequality of Lemma 7(a) with p replaced by n and putting $u = (C(\nu+2)n^{-1}\log p)^{1/2}$ yield

$$\mathbb{P}\left(\left\|n^{-1}\mathbf{E}'\mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max} > u\right) \le p^2 \max_{j,\ell}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\left(e_{ij}e_{\boldsymbol{\eta}^0,i\ell}\right)\right| > u\right) \le 2p^{-\nu}.$$

Finally, $W_{3,2}$ is found to have the same bound as $W_{1,3}$ because $\mathbf{E}_{\boldsymbol{\eta}^0}$ is an independent copy of $\mathbf{E}$. Consequently, with probability at least $1 - O(p^{-\nu})$, we obtain

$$\left\|\mathbf{U}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\right\|_{\max} \lesssim (n^{-1}\log p)^{1/2}.$$

This completes the proof of (a) since $p^{-\nu}/\pi_{np} = O(1)$.

Next we show (b) by verifying the following

$$\begin{aligned} &(b\text{–}i) \quad \sup_{|\mathcal{A}|\le k,\, \boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} \lesssim s^{3/2}\tilde{c}_{np}, \\ &(b\text{–}ii) \quad \left\|\mathbf{v}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{v}(\boldsymbol{\theta}^0)]\right\|_{\max} \lesssim s(n^{-1}\log p)^{1/2}. \end{aligned}$$

Similar to the proof of (a), we need to modify the proof of Lemma 8(b) in Section C.3 for obtaining the uniform bound with respect to $\mathcal{A}$, but the obtained result is already uniform over the choice of $\mathcal{A}$. Thus the same upper bound holds and $(b\text{–}i)$ follows. Next we show $(b\text{–}ii)$. It holds that

$$\begin{aligned} &\left\|\mathbf{v}(\boldsymbol{\theta}^0) - \mathbb{E}\,\mathbf{v}(\boldsymbol{\theta}^0)\right\|_{\max} \\ &\le \left\|n^{-1}\mathbf{X}'\mathbf{y} - \mathbb{E}\,n^{-1}\mathbf{X}'\mathbf{y}\right\|_{\max} + \left\|n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{y} - \mathbb{E}\,n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{y}\right\|_{\max} \\ &\le \left\|\left(n^{-1}\mathbf{X}'\mathbf{X} - \mathbb{E}\,n^{-1}\mathbf{X}'\mathbf{X}\right)\boldsymbol{\beta}\right\|_{\max} + \left\|n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \mathbb{E}\,n^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\right\|_{\max} \\ &+ \left\|\left(n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X} - \mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X}]\right)\boldsymbol{\beta}\right\|_{\max} + \left\|n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\boldsymbol{\varepsilon} - \mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\boldsymbol{\varepsilon}]\right\|_{\max} \\ &=: Z_1 + Z_2 + Z_3 + Z_4. \end{aligned}$$

These terms can be bounded by the results obtained in the proof of (a–ii). We see that

$$Z_1 \le s^{1/2} \left\| n^{-1}\mathbf{X}'\mathbf{X}_{\mathcal{S}^0} - \mathbb{E}\, n^{-1}\mathbf{X}'\mathbf{X}_{\mathcal{S}^0} \right\|_{\max} \|\boldsymbol{\beta}_{\mathcal{S}^0}\|_2 \lesssim sW_1 \lesssim s(n^{-1}\log p)^{1/2}$$

with probability at least $1 - O(p^{-\nu})$. Next we deduce

$$Z_2 \le \left\| n^{-1}\boldsymbol{\Lambda}^0\mathbf{F}^{0\prime}\boldsymbol{\varepsilon} \right\|_{\max} + \left\| n^{-1}\mathbf{E}'\boldsymbol{\varepsilon} \right\|_{\max}.$$

The first and second terms can be bounded by the same ways as $W_{1,3}$ and $W_{3,1}$ in the proof of (a) above with $\mathbf{E}$ and $\mathbf{E}_{\boldsymbol{\eta}^0}$ replaced by $\boldsymbol{\varepsilon}$, respectively. Then the first term dominates the second and hence $Z_2 \lesssim (n^{-1}\log p)^{1/2}$ with probability at least $1 - O(p^{-\nu})$. Similarly, we can obtain

$$Z_3 \le s^{1/2} \left\| n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X}_{\mathcal{S}^0} - \mathbb{E}\, n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X}_{\mathcal{S}^0} \right\|_{\max} \|\boldsymbol{\beta}_{\mathcal{S}^0}\|_2 \lesssim sW_3 \lesssim s(n^{-1}\log p)^{1/2}$$

with probability at least $1 - O(p^{-\nu})$. Note that Z_4 has the same bound as Z_2. Consequently, collecting terms leads to the result, $Z_1 + \cdots + Z_4 \lesssim s(n^{-1}\log p)^{1/2}$ with probability at least $1 - O(p^{-\nu})$. This proves (b–ii) and concludes the proof of Lemma 4.

B.3 Lemma 5 and its proof

Lemma 5 *Assume that all the conditions of Theorem 2 hold. Then with probability at least $1 - O(\pi_{np})$, the Lasso solution in (A.7) satisfies*

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| \hat{\boldsymbol{\beta}}^{aug}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{aug} \right\|_2 = O(s^{1/2}\lambda),$$

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| \hat{\boldsymbol{\beta}}^{aug}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{aug} \right\|_1 = O(s\lambda),$$

where $\lambda = c_1 n^{1/2}\log p$ with c_1 some positive constant.

Proof. Let $\boldsymbol{\delta}(:= \boldsymbol{\delta}(\boldsymbol{\theta})) := \hat{\boldsymbol{\beta}}^{\mathrm{aug}}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{\mathrm{aug}}$. We start with introducing two inequalities

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| n^{-1}[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]'\boldsymbol{\varepsilon} \right\|_{\max} \le 2^{-1}\lambda, \tag{B.7}$$

$$\inf_{\boldsymbol{\theta}\in\Theta_{np},\, \boldsymbol{\delta}\in\mathbb{V}} \boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \ge \sigma_e^2(1 + o(1)), \tag{B.8}$$

where $\lambda = c_1 n^{-1/2}\log p$ for some positive constant c_1 and

$$\mathbb{V} = \left\{ \boldsymbol{\delta} \in \mathbb{R}^{2p} : \|\boldsymbol{\delta}_{\mathcal{S}^1}\|_1 \le 3\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1,\ \|\boldsymbol{\delta}\|_0 \le k \right\}. \tag{B.9}$$

It is well known that the rate of convergence of the Lasso estimator can be obtained provided that (B.7) and (B.8) hold. Thus we show that these two inequalities actually hold with high probability in Step 1, and then derive the convergence rate using (B.7) and (B.8) in Step 2.

Step 1. We check whether (B.7) and (B.8) actually hold with high probability. We first verify (B.7). By the proofs of Lemmas 8 and 4, we have

$$\begin{aligned} &\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| n^{-1}[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]'\boldsymbol{\varepsilon} \right\|_{\max} \\ &\quad \le \left\| n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \right\|_{\max} + \sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta})'\boldsymbol{\varepsilon} - n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\boldsymbol{\varepsilon} \right\|_{\max} + \left\| n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\boldsymbol{\varepsilon} \right\|_{\max}. \end{aligned}$$

The first and third terms can both be upper bounded by $O(n^{-1/2}\log p)$ with probability at least $1-O(p^{-\nu})$, following the same lines for deriving bound for Z_2 in the proof of Lemma 4. To evaluate the second term, we can use the argument about V_2 and its upper bound (B.16) in the proof of Lemma 8. That bound still holds with the same rate $O(n^{-1/2}\log p)$ even if we take $\mathcal{A}=\{1,\dots,p\}$. Thus we conclude that (B.7) is true for the given λ by choosing an appropriate positive large constant c_1, with probability at least $1-O(\pi_{np})$.

Next to verify (B.8), we derive the population lower bound first and then show that the difference is negligible. From the construction, we have

$$\begin{aligned}
\mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)] &= \mathbb{E}[n^{-1}\mathbf{X}'\mathbf{X}] = \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}+\sigma_e^2\mathbf{I}_p,\\
\mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X}] &= \mathbb{E}[n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)] = \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}.
\end{aligned}$$

Using these equations, we obtain the lower bound

$$\begin{aligned}
\inf_{\boldsymbol{\delta}\in\mathbb{V}}\boldsymbol{\delta}'\,\mathbb{E}\left[\mathbf{U}(\boldsymbol{\theta}^0)\right]\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 &= \inf_{\boldsymbol{\delta}\in\mathbb{V}}\boldsymbol{\delta}'\begin{pmatrix}\boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}+\sigma_e^2\mathbf{I}_p & \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}\\ \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} & \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}+\sigma_e^2\mathbf{I}_p\end{pmatrix}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2\\
&= \inf_{\boldsymbol{\delta}\in\mathbb{V}}\boldsymbol{\delta}'\left\{\begin{pmatrix}1&1\\1&1\end{pmatrix}\otimes\boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}+\sigma_e^2\mathbf{I}_{2p}\right\}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2\\
&\geq \sigma_e^2. \qquad\qquad (B.10)
\end{aligned}$$

Because $\boldsymbol{\delta}\in\mathbb{V}$ is sparse and satisfies $|\mathcal{B}|\leq k$ for $\mathcal{B}:=\operatorname{supp}(\boldsymbol{\delta})$, it holds that $\boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta}^0)\boldsymbol{\delta}=\boldsymbol{\delta}_{\mathcal{B}}'\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)\boldsymbol{\delta}_{\mathcal{B}}$ and $\boldsymbol{\delta}'\,\mathbb{E}\left[\mathbf{U}(\boldsymbol{\theta}^0)\right]\boldsymbol{\delta}=\boldsymbol{\delta}_{\mathcal{B}}'\,\mathbb{E}\left[\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)\right]\boldsymbol{\delta}_{\mathcal{B}}$. Hence from Lemma 4 together with the condition on dimensionality, we obtain

$$\begin{aligned}
\sup_{|\mathcal{B}|\leq k,\,\boldsymbol{\theta}\in\Theta_{np}}\left\|\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta})-\mathbb{E}[\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)]\right\|_{\max} &= O\left(k^{1/2}\tilde{c}_{np}\right)\\
&= o(s^{-1}) \qquad\qquad (B.11)
\end{aligned}$$

with probability at least $1-O(\pi_{np})$. Thus using (B.11), we have for any $\boldsymbol{\delta}\in\mathbb{V}$,

$$\begin{aligned}
\boldsymbol{\delta}'\,\mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\boldsymbol{\delta}-\boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta} &= \boldsymbol{\delta}_{\mathcal{B}}'\left\{\mathbb{E}[\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)]-\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta})\right\}\boldsymbol{\delta}_{\mathcal{B}}\\
&\leq \|\boldsymbol{\delta}\|_1^2\sup_{|\mathcal{B}|\leq k,\,\boldsymbol{\theta}\in\Theta_{np}}\left\|\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta})-\mathbb{E}[\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)]\right\|_{\max} = \left(\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1+\|\boldsymbol{\delta}_{\mathcal{S}^1}\|_1\right)^2 o(s^{-1})\\
&\lesssim \|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1^2 o(s^{-1}) \leq \|\boldsymbol{\delta}_{\mathcal{S}^0}\|_2^2 o(1) \leq \|\boldsymbol{\delta}\|_2^2 o(1).
\end{aligned}$$

Rearranging the terms with (B.10) yields

$$\inf_{\boldsymbol{\theta}\in\Theta_{np},\,\boldsymbol{\delta}\in\mathbb{V}}\boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \geq \inf_{\boldsymbol{\delta}\in\mathbb{V}}\boldsymbol{\delta}'\,\mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2-|o(1)| \geq \sigma_e^2-|o(1)|,$$

resulting in (B.8). In consequence, two inequalities (B.7) and (B.8) hold with probability at least $1-O(\pi_{np})$.

Step 2. This part is well known in the literature (e.g., [33]) so we briefly give the proof omitting the details. Because the objective function is given by

$$\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) = \arg\min_{\mathbf{b}\in\mathbb{R}^{2p}} n^{-1}\left\|\mathbf{y}-[\mathbf{X},\widetilde{\mathbf{X}}(\boldsymbol{\theta})]\mathbf{b}\right\|_2^2+\lambda\|\mathbf{b}\|_1,$$

the global optimality of the Lasso estimator implies

$$(2n)^{-1}\left\|\mathbf{y}-[\mathbf{X},\widetilde{\mathbf{X}}(\boldsymbol{\theta})]\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})\right\|_2^2+\lambda\left\|\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta})\right\|_1 \\ \leq (2n)^{-1}\left\|\mathbf{y}-[\mathbf{X},\widetilde{\mathbf{X}}(\boldsymbol{\theta})]\boldsymbol{\beta}^{\mathsf{aug}}\right\|_2^2+\lambda\left\|\boldsymbol{\beta}^{\mathsf{aug}}\right\|_1,$$

where the true parameter vector $\boldsymbol{\beta}^{\mathsf{aug}}$ was defined in the proof of Theorem 2. Note that $\sup_{\boldsymbol{\theta}\in\Theta_{np}}\|\boldsymbol{\delta}(\boldsymbol{\theta})\|_0\leq k$ by the assumption. Expanding the inequality and collecting terms with (B.7) yield

$$2^{-1}\boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta}\leq\left\|n^{-1}\boldsymbol{\varepsilon}'[\mathbf{X},\widetilde{\mathbf{X}}(\boldsymbol{\theta})]\right\|_{\max}\|\boldsymbol{\delta}\|_1+\lambda\|\boldsymbol{\delta}\|_1\leq(3/2)\lambda\|\boldsymbol{\delta}\|_1. \tag{B.12}$$

On the other hand, applying Lemma 1 of [33] to our model reveals that $\boldsymbol{\delta}\in\mathbb{V}$. Thus we can use (B.8), (B.12), and (B.9) to get

$$\|\boldsymbol{\delta}\|_2^2(\sigma_e^2+o(1))\leq 3\lambda\|\boldsymbol{\delta}\|_1=3\lambda\left(\|\boldsymbol{\delta}_{\mathcal{S}^1}\|_1+\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1\right)\leq 12\lambda\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1.$$

Since $|\mathcal{S}^0|=s$ and $\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1\leq s^{1/2}\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_2$, it holds that $\|\boldsymbol{\delta}\|_2\leq 12s^{1/2}\lambda/(\sigma_e^2+o(1))$. Since $\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_2\leq\|\boldsymbol{\delta}\|_2$, we obtain the desired bound $\|\boldsymbol{\delta}\|_1\leq 48s\lambda/(\sigma_e^2+o(1))$. This bound holds uniformly over $\boldsymbol{\theta}\in\Theta_{np}$, which completes the proof of Lemma 5.

C Additional technical lemmas and their proofs

C.1 Lemma 6 and its proof

Lemma 6 *Denote by $\mathbf{V}\in\mathbb{R}^{r\times r}$ a diagonal matrix with its entries the r largest eigenvalues of $(np)^{-1}\mathbf{X}\mathbf{X}'$ and define $\mathbf{H}=(\boldsymbol{\Lambda}^{0\prime}\boldsymbol{\Lambda}^0/p)(\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n)\mathbf{V}^{-1}$. Assume that Conditions 2–5 hold. Then $\|\mathbf{H}\|_2+\|\mathbf{H}^{-1}\|_2+\|\mathbf{V}\|_2+\|\mathbf{V}^{-1}\|_2$ is bounded from above by some constant with probability at least $1-O(p^{-\nu})$.*

Proof. Let $\lambda^k[\mathbf{A}]$ denote the kth largest eigenvalue of square matrix $\mathbf{A}$ throughout the proof. Because $\|\boldsymbol{\Lambda}^{0\prime}\boldsymbol{\Lambda}^0/p\|_2\leq M$ and

$$\|\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n\|_2\leq\|n^{-1/2}\mathbf{F}^0\|_2\|n^{-1/2}\hat{\mathbf{F}}\|_2 \\ \leq(rn)^{1/2}\|n^{-1/2}\mathbf{F}^0\|_{\max}\left(\lambda^1[n^{-1}\hat{\mathbf{F}}'\hat{\mathbf{F}}]\right)^{1/2}\leq r^{1/2}M$$

by Conditions 2–3, and $\hat{\mathbf{F}}'\hat{\mathbf{F}}/n=\mathbf{I}_r$, we have

$$\|\mathbf{H}\|_2\leq\left\|\boldsymbol{\Lambda}^{0\prime}\boldsymbol{\Lambda}^0/p\right\|_2\left\|\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n\right\|_2\|\mathbf{V}^{-1}\|_2\lesssim\left\|\mathbf{V}^{-1}\right\|_2,$$

where $\|\mathbf{V}^{-1}\|_2$ is equal to the reciprocal of the rth largest eigenvalue of $(np)^{-1}\mathbf{X}\mathbf{X}'$. Similarly, under Conditions 2–3 we also have

$$\left\|\mathbf{H}^{-1}\right\|_2\leq\|\mathbf{V}\|_2\left\|(\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n)^{-1}\right\|_2\left\|(\boldsymbol{\Lambda}^{0\prime}\boldsymbol{\Lambda}^0/p)^{-1}\right\|_2\lesssim\|\mathbf{V}\|_2\left\|(\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n)^{-1}\right\|_2,$$

where $\|\mathbf{V}\|_2$ is equal to the largest eigenvalue of $(np)^{-1}\mathbf{X}\mathbf{X}'$ and the inverse matrix in the upper bound is well defined by [2]. To see if $\|(\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n)^{-1}\|_2$ is bounded from above, it suffices to bound the minimum eigenvalue of $\mathbf{F}^{0\prime}\hat{\mathbf{F}}\hat{\mathbf{F}}'\mathbf{F}^0/n^2$ away from zero uniformly in n. Regarding r eigenvalues of the matrix, Sylvester's law of inertia (e.g., [31], Theorem 4.5.8) entails that all the r eigenvalues are positive for all n. Moreover, by Proposition 1 of [2] we know that the limiting matrix of $\hat{\mathbf{F}}'\mathbf{F}^0/n$ is nonsingular under Conditions 2 and 5. Therefore, we can conclude that $\liminf_{n\to\infty} \lambda^r[\mathbf{F}^{0\prime}\hat{\mathbf{F}}\hat{\mathbf{F}}'\mathbf{F}^0/n^2] > 0$ a.s., and hence $\|\mathbf{H}^{-1}\|_2 \lesssim \|\mathbf{V}\|_2$ follows.

To complete the proof, it is sufficient to show that the maximum and rth largest eigenvalues of $(np)^{-1}\mathbf{X}\mathbf{X}'$ are bounded from above and away from zero, respectively, for all large n and p. By the definition of the spectral norm and triangle inequality, we have

$$\begin{aligned}\left\{\lambda^1\left[(np)^{-1}\mathbf{X}\mathbf{X}'\right]\right\}^{1/2} &= \left\|(np)^{-1/2}\mathbf{X}\right\|_2 \le \left\|(np)^{-1/2}\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\right\|_2 + \left\|(np)^{-1/2}\mathbf{E}\right\|_2 \\ &\le \left\|n^{-1/2}\mathbf{F}^0\right\|_2\left\|p^{-1/2}\mathbf{\Lambda}^0\right\|_2 + \left\|(np)^{-1/2}\mathbf{E}\right\|_2 .\end{aligned}$$

By Conditions 2 and 3, the first term is a.s. bounded by a constant as discussed above. The second term is $O((n\wedge p)^{-1/2}) = o(1)$ with probability at least $1-2\exp(-|O(n\vee p)|)$ by Lemma 9(a) under Condition 4. Therefore, the largest eigenvalue of $(np)^{-1}\mathbf{X}\mathbf{X}'$ is bounded from above by some constant with probability at least $1-2\exp(-|O(n\vee p)|)$.

Next we bound the rth largest eigenvalue of $(np)^{-1}\mathbf{X}\mathbf{X}'$ away from zero. Since the matrix is symmetric, Weyl's inequality (e.g., [31], Theorem 4.3.1) yields

$$\begin{aligned}\lambda^r\left[(np)^{-1}\mathbf{X}\mathbf{X}'\right] &= \lambda^r\left[(np)^{-1}\left\{\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \left(\mathbf{E}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{E}'\right) + \mathbf{E}\mathbf{E}'\right\}\right] \\ &\ge \lambda^r\left[(np)^{-1}\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}\right] + \lambda^n\left[(np)^{-1}\left(\mathbf{E}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{E}'\right)\right] + \lambda^n\left[(np)^{-1}\mathbf{E}\mathbf{E}'\right].\end{aligned} \tag{B.13}$$

The third term of lower bound (B.13) is obviously nonnegative. For the first term of lower bound (B.13), let $\mathcal{V}$ denote a subspace of $\mathbb{R}^n$. Because $\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}$ is symmetric, the Courant–Fischer min-max Theorem (e.g., [31], Theorem 4.2.6) yields

$$\begin{aligned}\lambda^r\left[(np)^{-1}\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}\right] &= \max_{\mathcal{V}:\dim(\mathcal{V})=r}\ \min_{\mathbf{v}\in\mathcal{V}\setminus\{\mathbf{0}\}}\left\{(np)^{-1}\frac{\mathbf{v}'\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}\mathbf{v}}{\mathbf{v}'\mathbf{v}}\right\} \\ &\ge \max_{\mathcal{V}:\dim(\mathcal{V})=r}\ \min_{\mathbf{v}\in\mathcal{V}\setminus\{\mathbf{0}\}}\left(n^{-1}\frac{\mathbf{v}'\mathbf{F}^0\mathbf{F}^{0\prime}\mathbf{v}}{\mathbf{v}'\mathbf{v}}\right)\min_{\mathbf{F}^{0\prime}\mathbf{v}\in\mathbb{R}^r\setminus\{\mathbf{0}\}}\left(p^{-1}\frac{\mathbf{v}'\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}\mathbf{v}}{\mathbf{v}'\mathbf{F}^0\mathbf{F}^{0\prime}\mathbf{v}}\right) \\ &= \lambda^r\left[n^{-1}\mathbf{F}^0\mathbf{F}^{0\prime}\right]\lambda^r\left[p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\right] = \lambda^r\left[n^{-1}\mathbf{F}^{0\prime}\mathbf{F}^0\right]\lambda^r\left[p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\right] \\ &\ge \lambda^r\left[\mathbf{\Sigma}_f\right]\lambda^r\left[p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\right] - \left\|n^{-1}\mathbf{F}^{0\prime}\mathbf{F}^0 - \mathbf{\Sigma}_f\right\|_2 \\ &\ge \lambda^r\left[\mathbf{\Sigma}_f\right]\lambda^r\left[p^{-1}\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\right] - r\left\|n^{-1}\mathbf{F}^{0\prime}\mathbf{F}^0 - \mathbf{\Sigma}_f\right\|_{\max}.\end{aligned}$$

In this lower bound, the first term is bounded away from zero by Conditions 2–3. Meanwhile, to evaluate the second term we use Lemma 7(d) in Section C.2, which together with the union bound establishes

$$\begin{aligned}\mathbb{P}\left(\left\|n^{-1}\mathbf{F}^{0\prime}\mathbf{F}^0 - \mathbf{\Sigma}_f\right\|_{\max} > u\right) &\le r^2\max_{k,\ell\in\{1,\dots,r\}}\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n\left(f_{ik}^0 f_{i\ell}^0 - \mathbb{E}\, f_{ik}^0 f_{i\ell}^0\right)\right| > u\right) \\ &\le 2r^2\exp(-nu^2/C)\end{aligned}$$

for any $0 \le u \le c$. Thus the second one turns out to be $O((n^{-1}\log p)^{1/2}) = o(1)$ with probability at least $1 - O(p^{-\nu})$ once we set $u = (C\nu n^{-1}\log p)^{1/2}$ and assume $n^{-1}\log p \le c^2/(C\nu)$ without loss of generality. Therefore, the first term of lower bound (B.13) is bounded away from zero eventually. For the second term of (B.13), since the spectral norm gives the upper bound of the spectral radius we have

$$\begin{aligned}\left|\lambda^n\left[(np)^{-1}\left(\mathbf{E}\boldsymbol{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\boldsymbol{\Lambda}^{0\prime}\mathbf{E}'\right)\right]\right| &\le \left\|(np)^{-1}\left(\mathbf{E}\boldsymbol{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\boldsymbol{\Lambda}^{0\prime}\mathbf{E}'\right)\right\|_2\\ &\le 2\left\|(np)^{-1/2}\mathbf{E}\right\|_2\left\|p^{-1/2}\boldsymbol{\Lambda}^0\right\|_2\left\|n^{-1/2}\mathbf{F}^0\right\|_2\\ &= O\left((n\wedge p)^{-1/2}\right)O(1)O(1) = o(1),\end{aligned}$$

which holds with probability at least $1 - 2\exp(-|O(n \vee p)|)$ by Lemma 9(a) in Section C.4. As a consequence, the desired result holds with probability at least $1 - O(p^{-\nu})$ and this concludes the proof of Lemma 6.

C.2 Lemma 7 and its proof

Lemma 7 *Assume that Conditions 2–4 hold. Then there exist some positive constants c and C such that the following inequalities hold*

(a) For all $\ell, i \in \{1, \ldots, n\}$ and $0 \le u \le c$, we have

$$\mathbb{P}\left(\left|p^{-1}\sum_{j=1}^{p}(e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}])\right| > u\right) \le 2\exp\left(-pu^2/C\right).$$

(b) For all $k \in \{1, \ldots, r\}$, $j \in \{1, \ldots, p\}$, and $0 \le u \le c$, we have

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} f_{ik}^0 e_{ij}\right| > u\right) \le 2\exp\left(-nu^2/C\right).$$

(c) For all $k \in \{1, \ldots, r\}$, $i \in \{1, \ldots, n\}$, and $u \ge 0$, we have

$$\mathbb{P}\left(\left|p^{-1}\sum_{j=1}^{p} \lambda_{jk}^0 e_{ij}\right| > u\right) \le 2\exp\left(-pu^2/C\right).$$

(d) For all $k, \ell \in \{1, \ldots, r\}$ and $0 \le u \le c$, we have

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\left(f_{ik}^0 f_{i\ell}^0 - \mathbb{E}[f_{ik}^0 f_{i\ell}^0]\right)\right| > u\right) \le 2\exp\left(-nu^2/C\right).$$

Proof. (a) To obtain the first result, we rely on the Hanson–Wright inequality. Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)' \in \mathbb{R}^m$ denote a random vector whose components are independent copies of $e \sim \text{subG}(C_e^2)$. Then the inequality states that for any (nonrandom) matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$,

$$\mathbb{P}\left(\left|\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi}\right| > u\right) \le 2\exp\left\{-\widetilde{C}_H \min\left(\frac{u^2}{K^4\|\mathbf{A}\|_F^2}, \frac{u}{K^2\|\mathbf{A}\|_2}\right)\right\}, \tag{B.14}$$

where K is a positive constant such that $\sup_{k\geq 1} k^{-1/2}(\mathbb{E}\,|e|^k)^{1/k} \leq K$ and $\widetilde{C}_H$ is a positive constant. In our setting, we can take $K = 3C_e^2$ (e.g., Lemma 1.4 of [34]). Using this inequality, we first prove the case when $\ell = i$. If we set $m = p$ and $\mathbf{A} = \operatorname{diag}(p^{-1}, \ldots, p^{-1})$, then we have

$$\left|\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi}\right| = \left|p^{-1}\sum_{j=1}^{p}(\xi_j^2 - \mathbb{E}\,\xi_j^2)\right| \overset{d}{=} \left|p^{-1}\sum_{j=1}^{p}\left(e_{ij}^2 - \mathbb{E}[e_{ij}^2]\right)\right|$$

for all i. Moreover, we obtain $\|\mathbf{A}\|_F^2 = p^{-1}$ and $\|\mathbf{A}\|_2 = p^{-1}$ in this case. The assumed condition $0 < u \leq 9C_e^2 = K^2$ entails that $u^2/K^4 \leq u/K^2$ so the result follows from (B.14) with $\widetilde{C}_H$ replaced by $C_H = 81C_e^4/\widetilde{C}_H$.

Similarly, we prove the case when $\ell \neq i$. We set $m = p+1$ and $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_{p+1})$, where $\mathbf{a}_1 = (0, p^{-1}, \ldots, p^{-1})'$ and $\mathbf{a}_j = \mathbf{0}$ for $j = 2, \ldots, p+1$. That is, the entries of $\mathbf{A}$ are all zero except that the second to $(p+1)$th components in the first column vector are p^{-1}. Under this setting, we observe that

$$\left|\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi}\right| = \left|p^{-1}\sum_{j=2}^{p+1}\xi_1\xi_j\right| \overset{d}{=} \left|p^{-1}\sum_{j=1}^{p}e_{\ell j}e_{ij}\right|$$

for all $\ell \neq i$. Moreover, we obtain $\|\mathbf{A}\|_F^2 = \|\mathbf{A}\|_2 = p^{-1}$ in this case. Therefore, the same bound holds as in the case of $\ell = i$ from (B.14) again. Consequently, for any $0 \leq u \leq 9C_e^2$ we have

$$\mathbb{P}\left(\left|p^{-1}\sum_{j=1}^{p}\left(e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}]\right)\right| > u\right) \leq 2\exp\left(-pu^2/C_H\right).$$

(b) We prove the second assertion by Bernstein's inequality for the sum of a martingale difference sequence (e.g., Theorem 3.14 in [11]). Fix $k = 1$ and $j = 1$. Define $\mathcal{F}_{i-1}$ as the σ-field generated from $\{f_{\ell 1}^0 : \ell = i, i-1, \ldots\}$. Then $(f_{i1}^0 e_{i1}, \mathcal{F}_i)$ forms a martingale difference sequence because $\mathbb{E}\,|f_{i1}^0 e_{i1}| < \infty$ and $\mathbb{E}[f_{i1}^0 e_{i1}|\mathcal{F}_{i-1}] = 0$ under Conditions 2 and 4. Since the sub-Gaussianity of e_{i1} implies $\mathbb{E}\,e_{i1}^2 \leq 4C_e^2$ (e.g., Lemma 1.4 of [34]), we have $V_i := \mathbb{E}\left[f_{ik}^{0\,2} e_{ij}^2 \mid \mathcal{F}_{i-1}\right] \leq 4C_e^2 M^2$, and hence $\sum_{i=1}^n V_i \leq 4nC_e^2M^2$ a.s. due to boundedness $|f_{i1}^0| \leq M$ a.s. On the other hand, by the sub-Gaussianity of e_{ij} and boundedness of $|f_{i1}^0|$ again we observe that for all $p \geq 3$ and $i \in \{1, \ldots, n\}$,

$$\mathbb{E}\left[(0 \vee f_{i1}^0 e_{i1})^p \mid \mathcal{F}_{i-1}\right] \leq M^p(2C_e^2)^{p/2}p\Gamma(p/2) \leq p!(2C_eM)^{p-2}V_i/2,$$

where Γ denotes the Gamma function and we have used the estimates $p\Gamma(p/2) \leq p!$ and $2^{p/2-2} \leq 2^{p-2}/2$ for $p \geq 3$ in the last inequality. Then an application of Theorem 3.14 in [11] by putting $x = u$, $y = 4M^2C_e^2$, and $c = 2MC_e$ in their notation gives the one-sided result. Making twice the bound yields

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}f_{ik}^0 e_{ij}\right| > u\right) \leq 2\exp\left(-\frac{nu^2}{8M^2C_e^2 + 4MC_eu}\right).$$

For all $0 \le u \le MC_e^2$, the upper bound is further bounded by $2\exp(nu^2/(12M^2C_e^2))$. We set $C_I = 12M^2C_e^2$. Consequently, for any $0 \le u \le MC_e^2$ we have

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} f_{ik}^0 e_{ij}\right| > u\right) \le 2\exp\left(-nu^2/C_I\right).$$

(c) We prove the third inequality. Note that

$$\mathbb{P}\left(\left|\lambda_{jk}^0 e_{ij}\right| > u\right) \le 2\exp\left\{-\frac{u^2}{2\lambda_{jk}^{02}C_e^2}\right\} \le 2\exp\left\{-\frac{u^2}{2M^2C_e^2}\right\}.$$

This implies that $\lambda_{jk}^0 e_{ij}$ is a sequence of i.i.d. subG$(M^2C_e^2)$. Thus the result is obtained directly by Bernstein's inequality for the sum of independent sub-Gaussian random variables. Consequently, for any $u \ge 0$ putting $C_J = M^2C_e^2$ leads to

$$\mathbb{P}\left(\left|p^{-1}\sum_{j=1}^{p} \lambda_{jk}^0 e_{ij}\right| > u\right) \le 2\exp\left(-pu^2/C_J\right).$$

(d) We show the last inequality. Note that for each k, $(f_{ik})_i \sim$ i.i.d. subG(M^2) since $|f_{ik}^0| \le M$ a.s. by Lemma 1.8 of [34] under Condition 2. Thus the remaining is the same as (a). Set $C_K = 81M^4/\widetilde{C}_H$ here. Then for any $0 \le u \le 9M^2$, we have

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}\left(f_{ik}^0 f_{i\ell}^0 - \mathbb{E}[f_{ik}^0 f_{i\ell}^0]\right)\right| > u\right) \le 2\exp\left(-nu^2/C_K\right).$$

Finally the obtained inequalities hold even if the constant in the upper bound is replaced with arbitrary fixed constant C such that $C \ge \max\{C_H, C_I, C_J, C_K\}$. Similarly, we can also restrict the range of u for each inequality to be $0 \le u \le c$ for arbitrary fixed constant c that satisfies $0 < c \le \min(9C_e^2, MC_e^2, 9M^2)$. This completes the proof of Lemma 7.

C.3 Lemma 8 and its proof

Lemma 8 *Assume that Conditions 1–4 hold. Then for any set $\mathcal{A}$ satisfying $|\mathcal{A}| \le k$, the following statements hold with probability at least $1 - O(\pi_{np})$*

$$(a) \quad \sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} = O\left(k^{1/2}\tilde{c}_{np}\right),$$

$$(b) \quad \sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} = O\left(s^{3/2}\tilde{c}_{np}\right),$$

where Θ_{np} was defined in Lemma 3 and $\tilde{c}_{np} = n^{-1/2}\log p + p^{-1/2}\log n$. Consequently, we have

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} = O\left(\left(k^{1/2} + s^{3/2}\right)\tilde{c}_{np}\right).$$

Proof. We first state some results that are useful in the proof. Since $\|n^{-1/2}\mathbf{F}^0\|_2 = O(1)$ a.s. by Condition 2 and $\|k^{-1/2}\boldsymbol{\Lambda}_{\mathcal{A}}^0\|_2 = O(1)$ for any $\mathcal{A}$ such that $|\mathcal{A}| \le k$ under Condition 3, we first have

$$\left\|n^{-1/2}\mathbf{C}_{\mathcal{A}}^0\right\|_2 \le \left\|n^{-1/2}\mathbf{F}^0\right\|_2 k^{1/2}\left\|k^{-1/2}\boldsymbol{\Lambda}_{\mathcal{A}}^0\right\|_2 \lesssim k^{1/2}.$$

Next Lemma 9(b) in Section C.4 gives directly

$$\left\| n^{-1/2}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_2 \lesssim 1 \tag{B.15}$$

with probability at least $1 - O(p^{-\nu})$. By Condition 4, we also deduce

$$\begin{aligned}
\mathbb{P}\left(\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}_{np}} \left\|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max} > u \right) &\le np \max_{i,j} \mathbb{P}\left(\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}_{np}} \left|e_{\boldsymbol{\eta}ij} - e_{\boldsymbol{\eta}^0ij}\right| > u \right) \\
&\le np \max_{i,j} \mathbb{P}\left(|Z| > u/(M^{1/2}c_{np}^{1/2}) \right) \\
&\le 2np \exp\left(-u^2/\left(c_e^2 M c_{np}\right)\right)
\end{aligned}$$

for any $u \ge 0$. Thus setting $u = 2c_e M^{1/2} c_{np}^{1/2} \log^{1/2}(np)$ with some large enough positive constant M, we obtain that with probability at least $1 - O((np)^{-\nu})$,

$$\sup_{\boldsymbol{\eta}\in\boldsymbol{\Theta}_{np}} \left\|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max} \lesssim c_{np}\log^{1/2}(np) = O(\tilde{c}_{np}).$$

We will use these results and Lemma 10 in Section C.5 in the proofs below.

To prove (a), we have

$$\begin{aligned}
\left\|\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max} &\le \left\| n^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})'\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}) - n^{-1}\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0) \right\|_{\max} \\
&\quad + 2\left\| n^{-1}\mathbf{X}_{\mathcal{A}}'\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}) - n^{-1}\mathbf{X}_{\mathcal{A}}'\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0) \right\|_{\max} =: U_1 + U_2.
\end{aligned}$$

Observe that U_1 is further bounded as

$$\begin{aligned}
U_1 &\le \left\| n^{-1}\mathbf{C}_{\mathcal{A}}'\mathbf{C}_{\mathcal{A}} - n^{-1}\mathbf{C}_{\mathcal{A}}^{0\,\prime}\mathbf{C}_{\mathcal{A}}^0 \right\|_{\max} + \left\| n^{-1}\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}}'\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - n^{-1}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}'\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_{\max} \\
&\quad + 2\left\| n^{-1}\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}}'\mathbf{C}_{\mathcal{A}} - n^{-1}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}'\mathbf{C}_{\mathcal{A}}^0 \right\|_{\max} =: U_{11} + U_{12} + U_{13}.
\end{aligned}$$

By Lemma 10, it is easy to see that

$$\begin{aligned}
U_{11} &\le \left\| n^{-1}\left(\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right)'\left(\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right) \right\|_{\max} + 2\left\| n^{-1}\mathbf{C}_{\mathcal{A}}^{0\,\prime}\left(\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right) \right\|_{\max} \\
&\le n^{-1/2}\left\|\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right\|_{\max}\left\|\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right\|_2 + 2\left\| n^{-1/2}\mathbf{C}_{\mathcal{A}}^0 \right\|_2 \left\|\mathbf{C}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^0\right\|_{\max} \\
&\lesssim k^{1/2}\left\|\mathbf{C} - \mathbf{C}^0\right\|_{\max}^2 + k^{1/2}\left\|\mathbf{C} - \mathbf{C}^0\right\|_{\max} \\
&= O\left(k^{1/2}c_{np}^2 + k^{1/2}c_{np}\right) = O\left(k^{1/2}c_{np}\right),
\end{aligned}$$

where the last estimate follows from Lemma 3. Similarly, we deduce

$$\begin{aligned}
U_{12} &\le \left\| n^{-1}\left(\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right)'\left(\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right) \right\|_{\max} + 2\left\| n^{-1}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}'\left(\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right) \right\|_{\max} \\
&\le n^{-1/2}\left\|\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right\|_{\max}\left\|\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right\|_2 + 2\left\| n^{-1/2}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_2 \left\|\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\right\|_{\max} \\
&\lesssim k^{1/2}\left\|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max}^2 + \left\|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max} \\
&= O\left(k^{1/2}\tilde{c}_{np}^2 + \tilde{c}_{np}\right)
\end{aligned}$$

and

$$\begin{aligned}
U_{13} &\leq \left\| n^{-1} \left(\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right)' \left(\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}} \right) \right\|_{\max} \\
&\quad + \left\| n^{-1} \mathbf{E}'_{\boldsymbol{\eta}^0\mathcal{A}} \left(\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}} \right) \right\|_{\max} + \left\| n^{-1} \mathbf{C}^{0\,\prime}_{\mathcal{A}} \left(\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}^0_{\boldsymbol{\eta}^0\mathcal{A}} \right) \right\|_{\max} \\
&\leq k^{1/2} \left\| \mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0} \right\|_{\max} \left\| \mathbf{C} - \mathbf{C}^0 \right\|_{\max} \\
&\quad + \left\| n^{-1/2} \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_2 \left\| \mathbf{C} - \mathbf{C}^0 \right\|_{\max} + \left\| n^{-1/2} \mathbf{C}^0_{\mathcal{A}} \right\|_2 \left\| \mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}^0_{\boldsymbol{\eta}^0} \right\|_{\max} \\
&= O\left(k^{1/2} \tilde{c}_{np} c_{np} + c_{np} + k^{1/2} \tilde{c}_{np} \right) = O\left(k^{1/2} \tilde{c}_{np} \right).
\end{aligned}$$

Combining these bounds of U_{11}–U_{13}, we have

$$U_1 \leq U_{11} + U_{12} + U_{13} \lesssim k^{1/2} \tilde{c}_{np}.$$

This holds uniformly in $\boldsymbol{\theta} \in \Theta_{np}$ with probability at least $1 - O(\pi_{np})$ by Lemma 3 and the discussion above. Next we obtain

$$\begin{aligned}
U_2 &\leq \left\| n^{-1} \mathbf{C}^{0\,\prime}_{\mathcal{A}} (\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}}) \right\|_{\max} + \left\| n^{-1} \mathbf{C}^{0\,\prime}_{\mathcal{A}} (\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}) \right\|_{\max} \\
&\quad + \left\| n^{-1} \mathbf{E}'_{\mathcal{A}} (\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}}) \right\|_{\max} + \left\| n^{-1} \mathbf{E}'_{\mathcal{A}} (\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}) \right\|_{\max} \\
&\leq \left\| n^{-1/2} \mathbf{C}^0_{\mathcal{A}} \right\|_2 \left\| \mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}} \right\|_{\max} + \left\| n^{-1/2} \mathbf{C}^0_{\mathcal{A}} \right\|_2 \left\| \mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_{\max} \\
&\quad + \left\| n^{-1/2} \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_2 \left\| \mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}} \right\|_{\max} + \left\| n^{-1/2} \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}} \right\|_2 \left\| \mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0} \right\|_{\max} \\
&= O\left(k^{1/2} c_{np} + k^{1/2} \tilde{c}_{np} + c_{np} + \tilde{c}_{np} \right) \\
&= O\left(k^{1/2} \tilde{c}_{np} \right).
\end{aligned}$$

This also holds uniformly in $\boldsymbol{\theta} \in \Theta_{np}$ with probability at least $1 - O(\pi_{np})$ by Lemma 3 and the discussion above. Consequently, it holds that

$$\sup_{\boldsymbol{\theta} \in \Theta_{np}} \left\| \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}^0) \right\|_{\max} \lesssim k^{1/2} \tilde{c}_{np}$$

with probability at least $1 - O(\pi_{np})$.

To prove (b), we have

$$\begin{aligned}
&\left\| \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}^0) \right\|_{\max} \leq \left\| n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})' \mathbf{y} - n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)' \mathbf{y} \right\|_{\max} \\
&\quad \leq \left\| n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})' \mathbf{X}\boldsymbol{\beta} - n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)' \mathbf{X}\boldsymbol{\beta} \right\|_{\max} + \left\| n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})' \boldsymbol{\varepsilon} - n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)' \boldsymbol{\varepsilon} \right\|_{\max} \\
&\quad =: V_1 + V_2.
\end{aligned}$$

First, because $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_{\mathcal{S}^0} \boldsymbol{\beta}_{\mathcal{S}^0}$ we see that

$$\begin{aligned}
V_1 &\leq s^{1/2} \left\| n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})' \mathbf{X}_{\mathcal{S}^0} - n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)' \mathbf{X}_{\mathcal{S}^0} \right\|_{\max} \left\| \boldsymbol{\beta}_{\mathcal{S}^0} \right\|_2 \\
&\lesssim s \left\| n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})' \mathbf{X}_{\mathcal{S}^0} - n^{-1} \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)' \mathbf{X}_{\mathcal{S}^0} \right\|_{\max}.
\end{aligned}$$

Recall that $|\mathcal{S}^0| = s$ and $s \le n \wedge p$. By a similar bound of U_2, the norm just above can be bounded further as

$$\begin{aligned}
&\left\| n^{-1/2}\mathbf{C}^0_{\mathcal{S}^0} \right\|_2 \|\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}}\|_{\max} + \left\| n^{-1/2}\mathbf{C}^0_{\mathcal{S}^0} \right\|_2 \|\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\|_{\max} \\
&\qquad + \left\| n^{-1/2}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{S}^0} \right\|_2 \|\mathbf{C}_{\mathcal{A}} - \mathbf{C}^0_{\mathcal{A}}\|_{\max} + \left\| n^{-1/2}\mathbf{E}_{\boldsymbol{\eta}^0\mathcal{S}^0} \right\|_2 \|\mathbf{E}_{\boldsymbol{\eta}\mathcal{A}} - \mathbf{E}_{\boldsymbol{\eta}^0\mathcal{A}}\|_{\max} \\
&\lesssim s^{1/2} \|\mathbf{C} - \mathbf{C}^0\|_{\max} + s^{1/2} \|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\|_{\max} + \|\mathbf{C} - \mathbf{C}^0\|_{\max} + \|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\|_{\max} \\
&= O\left(s^{1/2} c_{np} + s^{1/2}\tilde{c}_{np} + c_{np} + \tilde{c}_{np}\right) = O\left(s^{1/2}\tilde{c}_{np}\right).
\end{aligned}$$

Thus we have

$$V_1 \lesssim s s^{1/2} \tilde{c}_{np} = s^{3/2}\tilde{c}_{np}$$

with probability at least $1 - O(\pi_{np})$. Next the same procedure yields

$$\begin{aligned}
V_2 &\le \left\| \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}) - \widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0) \right\|_{\max} \left\| n^{1/2}\boldsymbol{\varepsilon} \right\|_2 \\
&\lesssim \|\mathbf{C} - \mathbf{C}^0\|_{\max} + \|\mathbf{E}_{\boldsymbol{\eta}} - \mathbf{E}_{\boldsymbol{\eta}^0}\|_{\max} \lesssim \tilde{c}_{np},
\end{aligned} \tag{B.16}$$

where $\|n^{1/2}\boldsymbol{\varepsilon}\|_2 = (\mathbb{E}\,\varepsilon^2)^{1/2} + o(1)$ a.s. by the law of large numbers for independent random variables. Since the results hold uniformly in $\boldsymbol{\theta} \in \Theta_{np}$, combining them leads to

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \|\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - \mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}^0)\|_{\max} \lesssim s^{3/2}\tilde{c}_{np}$$

with probability at least $1 - O(\pi_{np})$. This concludes the proof of Lemma 8.

C.4 Lemma 9 and its proof

Lemma 9 *Assume that Condition 4 holds. Then the following statements hold*

(a) We have

$$\mathbb{P}\left(\left\| (n \vee p)^{-1/2}\mathbf{E} \right\|_2 \lesssim 1 \right) \ge 1 - 2\exp(-|O(n \vee p)|);$$

(b) For any fixed set $\mathcal{A}$ with $|\mathcal{A}| \le k \le n$, we have

$$\mathbb{P}\left(\left\| n^{-1/2}\mathbf{E}_{\mathcal{A}} \right\|_2 \lesssim 1 \right) \ge 1 - 2p^{-\nu};$$

(c) For all $k \le n$, we have

$$\mathbb{P}\left(\max_{|\mathcal{A}| \le k} \left\| n^{-1/2}\mathbf{E}_{\mathcal{A}} \right\|_2 \lesssim 1 \vee \left(n^{-1} k \log p\right)^{1/2} \right) \ge 1 - 2p^{-\nu},$$

where $\nu > 0$ is a predetermined constant.

Proof. Result (a) is obtained by Theorem 5.39 of [40]. Moreover, by the same theorem there exist some positive constants c and C such that for any $\mathcal{A}$ with $|\mathcal{A}| \le k \le n$ and every $t \ge 0$,

$$\mathbb{P}\left(\sigma_e^{-1} \|n^{-1/2}\mathbf{E}_{\mathcal{A}}\|_2 > 1 + C + n^{-1/2}t \right) \le 2\exp\left(-ct^2\right),$$

where $\sigma_e^2 = \mathbb{E}\, e^2$. Therefore, result (b) is immediately obtained by putting $t^2 = c^{-1}\nu \log p$ since $n^{-1/2}t = o(1)$ and $\exp\left(-ct^2\right) = p^{-\nu}$ in this case.

For (c), taking the union bound leads to

$$\begin{aligned}&\mathbb{P}\left(\sigma_e^{-1} \max_{|\mathcal{A}|\le k} \|n^{-1/2}\mathbf{E}_{\mathcal{A}}\|_2 > 1 + C + n^{-1/2}t\right)\\ &\quad\le \binom{p}{k} \max_{|\mathcal{A}|\le k} \mathbb{P}\left(\sigma_e^{-1}\|n^{-1/2}\mathbf{E}_{\mathcal{A}}\|_2 > 1 + C + n^{-1/2}t\right) \le 2p^k \exp\left(-ct^2\right).\end{aligned}$$

Set $t^2 = c^{-1}(\nu + k)\log p$ in this inequality. Then we have $n^{-1/2}t = O\left((n^{-1}k\log p)^{1/2}\right)$ and

$$2p^k \exp\left(-ct^2\right) \le 2p^k \exp\left(-(\nu + k)\log p\right) = 2p^{-\nu},$$

which gives result (c) and completes the proof of Lemma 9.

C.5 Lemma 10 and its proof

Lemma 10 *For matrices* $\mathbf{A} \in \mathbb{R}^{k_1\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{n\times k_2}$*, we have* $\|\mathbf{AB}\|_{\max} \le n^{1/2}\|\mathbf{A}\|_2\|\mathbf{B}\|_{\max}$ *and* $\|\mathbf{AB}\|_{\max} \le n^{1/2}\|\mathbf{A}\|_{\max}\|\mathbf{B}\|_2$.

Proof. For any matrix $\mathbf{M} = (m_{ij}) \in \mathbb{R}^{k\times n}$, let $\|\mathbf{M}\|_{\infty,\infty}$ denote the induced ℓ_∞-norm. First, we have

$$\|\mathbf{M}\|_{\infty,\infty} := \sup_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Mv}\|_{\max}}{\|\mathbf{v}\|_{\max}} \le \sup_{\mathbf{v}\in\mathbb{R}^n\setminus\{\mathbf{0}\}} \frac{\|\mathbf{Mv}\|_2}{\|\mathbf{v}\|_2}\frac{\|\mathbf{v}\|_2}{\|\mathbf{v}\|_{\max}} \le n^{1/2}\|\mathbf{M}\|_2.$$

Therefore, by a simple calculation we see that

$$\begin{aligned}\|\mathbf{AB}\|_{\max} &= \|\operatorname{vec}(\mathbf{AB})\|_{\max} = \|(\mathbf{I}_{k_2}\otimes\mathbf{A})\operatorname{vec}(\mathbf{B})\|_{\max}\\ &= \frac{\|(\mathbf{I}_{k_2}\otimes\mathbf{A})\operatorname{vec}(\mathbf{B})\|_{\max}}{\|\operatorname{vec}(\mathbf{B})\|_{\max}}\|\operatorname{vec}(\mathbf{B})\|_{\max}\\ &\le \|\mathbf{I}_{k_2}\otimes\mathbf{A}\|_{\infty,\infty}\|\operatorname{vec}(\mathbf{B})\|_{\max} = \|\mathbf{A}\|_{\infty,\infty}\|\mathbf{B}\|_{\max} \le n^{1/2}\|\mathbf{A}\|_2\|\mathbf{B}\|_{\max}.\end{aligned}$$

The second assertion follows from applying this inequality to $\mathbf{B}'\mathbf{A}'$. This concludes the proof of Lemma 10.