Supplementary Material to "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]

August 5, 2019

This Supplementary Material contains a review of the model-X knockoffs framework, the proofs of Proposition 1 and Lemmas 1–2, and additional technical details and numerical results. All the notation is the same as in the main body of the paper.

B 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 concept was introduced originally in [8] and whose definition is stated formally as follows for completeness.

Definition 1 (Model-X knockoff variables [8]) For a set of random variables $\mathbf{x} = (X_1, \cdots, X_p)$, 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:

*The author names are alphabetically ordered. 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. Most of this work was completed while Uematsu visited USC as a JSPS Overseas Research Fellow and Postdoctoral Scholar. The authors sincerely thank the Joint Editor, Associate Editor, and referees for their valuable comments that helped improve the paper substantially.

1) For any subset $\mathcal{S} \subset \{1, \cdots, p\}$, we have $[\mathbf{x}, \widetilde{\mathbf{x}}]_{\mathsf{swap}(\mathcal{S})} \stackrel{d}{=} [\mathbf{x}, \widetilde{\mathbf{x}}]$, where $\stackrel{d}{=}$ denotes equal in distribution and 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 $\mathcal{S} \subset \mathcal{S}^1$,

$$([\mathbf{x}, \widetilde{\mathbf{x}}]_{\mathsf{swap}(\mathcal{S})}, Y) \stackrel{d}{=} ([\mathbf{x}, \widetilde{\mathbf{x}}], Y). \tag{A.1}$$

Such a property is known as the *exchangeability property* using the terminology in [8]. For more details, see Lemma 3.2 therein. Following [8], 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, [8] suggested constructing knockoff statistics $W_j = w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y})$, $j \in \{1, \cdots, p\}$, 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,

$$w_j([\mathbf{X}, \widetilde{\mathbf{X}}]_{\mathsf{swap}(\mathcal{S})}, \mathbf{y}) = \begin{cases} w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}), & j \notin \mathcal{S}, \\ -w_j([\mathbf{X}, \widetilde{\mathbf{X}}], \mathbf{y}), & j \in \mathcal{S}. \end{cases} \tag{A.2}$$

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 [8].

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

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

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, [8] proposed to choose the threshold by resorting to the above $\widehat{\mathrm{FDP}}$. Their results are summarized formally as follows.

Result 1 ([8]) *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, [5] 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. [10] 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 [8], [10], and [5].

C Proofs of Proposition 1 and some key lemmas

C.1 Proof of Proposition 1

Observe that the second property of Definition 1 holds naturally since $\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)$ is constructed without using the information of $\mathbf{y}$. Thus it remains to verify the first property of Definition 1. Since $\mathbf{F}^0$ and $\mathbf{E}_{\boldsymbol{\eta}^0}$ have i.i.d. rows, let us consider the case of a single observation and show that $[\mathbf{x}, \widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)]_{\mathsf{swap}(\mathcal{S})} \overset{d}{=} [\mathbf{x}, \widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)]$ for any subset $\mathcal{S} \subset \{1, \cdots, p\}$. By Proposition 2 of [8], it suffices to consider the case of $\mathcal{S} = \{j\}$ for an arbitrary $j \in \{1, \cdots, p\}$. It follows from the definition of model (2) and the construction of $\widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)$ that

$$\begin{aligned}[\mathbf{x}, \widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)]_{\mathsf{swap}(\{j\})} &= [\mathbf{c}^0 + \mathbf{e}, \mathbf{c}^0 + \mathbf{e}_{\boldsymbol{\eta}^0}]_{\mathsf{swap}(\{j\})} \\ &= [\mathbf{c}^0 + \widetilde{\mathbf{e}}^{(j)}, \mathbf{c}^0 + \widetilde{\mathbf{e}}^{(j)}_{\boldsymbol{\eta}^0}], \end{aligned} \tag{A.3}$$

where $\widetilde{\mathbf{e}}^{(j)}$ and $\widetilde{\mathbf{e}}^{(j)}_{\boldsymbol{\eta}^0}$ are defined such that $[\mathbf{e}, \mathbf{e}_{\boldsymbol{\eta}^0}]_{\mathsf{swap}(\{j\})} = [\widetilde{\mathbf{e}}^{(j)}, \widetilde{\mathbf{e}}^{(j)}_{\boldsymbol{\eta}^0}]$. Since model (2) assumes that $\mathbf{e}$ has i.i.d. components and $\mathbf{e}_{\boldsymbol{\eta}^0}$ is an independent copy of $\mathbf{e}$, it holds that

$$[\widetilde{\mathbf{e}}^{(j)}, \widetilde{\mathbf{e}}^{(j)}_{\boldsymbol{\eta}^0}] \stackrel{d}{=} [\mathbf{e}, \mathbf{e}_{\boldsymbol{\eta}^0}]. \tag{A.4}$$

Therefore, in view of (A.3) and (A.4) and the independence between $(\mathbf{e}, \mathbf{e}_{\boldsymbol{\eta}^0})$ and $\mathbf{c}^0$, we have

$$\begin{aligned}[\mathbf{x}, \widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)]_{\mathsf{swap}(\{j\})} &\stackrel{d}{=} [\mathbf{c}^0 + \mathbf{e}, \mathbf{c}^0 + \mathbf{e}_{\boldsymbol{\eta}^0}] \\ &= [\mathbf{x}, \widetilde{\mathbf{x}}(\boldsymbol{\theta}^0)],\end{aligned}$$

which completes the proof of Proposition 1.

C.2 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:

$$\mathbf{v}(\boldsymbol{\theta}) - \mathbf{U}(\boldsymbol{\theta})\hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) = n^{-1}\lambda \mathbf{z}, \tag{A.5}$$

$$\text{where } \mathbf{z} = (z_1, \cdots, z_{2p})^T \text{ with } z_j \in \begin{cases} \{\mathrm{sgn}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \hat{\beta}_j = 0, \end{cases} \quad \text{for} \quad j = 1, \cdots, 2p. \tag{A.6}$$

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.

C.3 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:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta})'\mathbf{U}_{\mathcal{A}}(\boldsymbol{\theta}))^{-1}(\mathbf{v}_{\mathcal{A}}(\boldsymbol{\theta}) - n^{-1}\lambda\tilde{\mathbf{z}}), \tag{A.7}$$

$$\text{where } \tilde{\mathbf{z}} = (\tilde{z}_1, \cdots, \tilde{z}_{2|\mathcal{A}|})^T \text{ with } \tilde{z}_j \in \begin{cases} \{\mathrm{sgn}(\tilde{\beta}_j)\} & \text{if } \tilde{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \tilde{\beta}_j = 0, \end{cases} \quad \text{for} \quad j = 1, \cdots, 2|\mathcal{A}|. \tag{A.8}$$

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.7)–(A.8) and the Lasso solution to (A.5)–(A.6) have the identical support (when viewed in the original $2p$-dimensional space) and in addition, identical values on the support. This guarantees that $S_{\{1,\cdots,p\}}(\mathbf{T}(\boldsymbol{\theta}))$ and $S_{\mathcal{A}}(\mathbf{T}_{\mathcal{A}}(\boldsymbol{\theta}))$ are identical and thus concludes the proof of Lemma 2.

C.4 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 $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, p\}$. 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 D.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 $\max_{i \in \{1, \cdots, n\}} \|\hat{\mathbf{f}}_i\|_2^2 \lesssim 1$ a.s. by the assumed restriction $\hat{\mathbf{F}}' \hat{\mathbf{F}} / n = \mathbf{I}_r$ as mentioned on p.213 of [2]. Hence, the triangle and Cauchy–Schwarz inequalities with Conditions 2 and 3 give

$$\begin{aligned} \max_{i,j} |\hat{c}_{ij} - c_{ij}| &\leq \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 \max_j \|\boldsymbol{\lambda}_j^*\|_2 + \max_i \|\hat{\mathbf{f}}_i\|_2 \max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2 \\ &\lesssim \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 + \max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2. \end{aligned} \tag{A.9}$$

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 [1] along with taking maximum

over $i, \ell \in \{1, \cdots, n\}$, we can deduce

$$
\begin{aligned}
&\max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 \\
&\quad \le \|\mathbf{V}^{-1}\|_2 \max_i \left((\sigma_e^2/n)\|\hat{\mathbf{f}}_i\|_2 + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell\|_2 \left| p^{-1}\sum_{j=1}^p (e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}]) \right| \right. \\
&\qquad \left. + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell \mathbf{f}_\ell^{0\prime}\|_2 \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{ij} \right\|_2 + n^{-1}\sum_{\ell=1}^n \|\hat{\mathbf{f}}_\ell \mathbf{f}_i^{0\prime}\|_2 \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{\ell j} \right\|_2 \right) \\
&\quad \lesssim O(n^{-1}) + \max_{i,\ell} \left| p^{-1}\sum_{j=1}^p (e_{\ell j}e_{ij} - \mathbb{E}[e_{\ell j}e_{ij}]) \right| + \max_i \left\| p^{-1}\sum_{j=1}^p \boldsymbol{\lambda}_j^0 e_{ij} \right\|_2 \\
&\quad \lesssim O(n^{-1}) + R_1 + R_2, \qquad \text{(A.10)}
\end{aligned}
$$

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 [1] with taking maximum over $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, p\}$ leads to

$$
\begin{aligned}
&\max_j \|\hat{\boldsymbol{\lambda}}_j - \boldsymbol{\lambda}_j^*\|_2 \\
&\quad \le \|\mathbf{H}\|_2 \max_j \left\| n^{-1}\sum_{i=1}^n \mathbf{f}_i^0 e_{ij} \right\|_2 + \left\| n^{-1}\sum_{i=1}^n \hat{\mathbf{f}}_i(\hat{\mathbf{f}}_i - \mathbf{f}_i^*)' \right\|_2 \left\| \mathbf{H}^{-1} \right\|_2 \max_j \left\| \boldsymbol{\lambda}_j^0 \right\|_2 \\
&\qquad + \max_j \left\| n^{-1}\sum_{i=1}^n (\hat{\mathbf{f}}_i - \mathbf{f}_i^*) e_{ij} \right\|_2 \\
&\quad \lesssim \max_j \left\| n^{-1}\sum_{i=1}^n \mathbf{f}_i^0 e_{ij} \right\|_2 + \max_i \left\| \hat{\mathbf{f}}_i - \mathbf{f}_i^* \right\|_2 + \max_i \left\| \hat{\mathbf{f}}_i - \mathbf{f}_i^* \right\|_2 \max_j \left(n^{-1}\sum_{i=1}^n e_{ij}^2 \right)^{1/2} \\
&\quad = R_3 + \max_i \|\hat{\mathbf{f}}_i - \mathbf{f}_i^*\|_2 (1 + R_4), \qquad \text{(A.11)}
\end{aligned}
$$

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 D.2 with p replaced by n and the union bound give

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

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 (A.9)–(A.11) yields

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

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 D.2 and the union bound with condition $c_{np} \le c/(r^2M^2C(\nu+2))^{1/2}$ as above. First for R_1, Lemma 7(a) by letting $u_1 = (C(\nu+2)p^{-1}\log n)^{1/2}$ results in

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

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

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

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

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

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 $k \in \{1, \cdots, m\}$, the binomial expansion entails

$$\begin{aligned}
\left| \mathbb{E}_{np} \hat{e}^k - \mathbb{E}\, e^k \right| &= \left| \mathbb{E}_{np} (e + \delta)^k - \mathbb{E}\, e^k \right| \\
&= \left| \mathbb{E}_{np} (e^k - \mathbb{E}\, e^k) + \mathbb{E}_{np} \sum_{\ell=0}^{k-1} \binom{k}{\ell} e^\ell \delta^{k-\ell} \right| \\
&\leq \left| \mathbb{E}_{np} (e^k - \mathbb{E}\, e^k) \right| + \sum_{\ell=0}^{k-1} \binom{k}{\ell} \max_{i,j} |\delta_{ij}|^{k-\ell} \mathbb{E}_{np} |e|^\ell \\
&\lesssim \left| \mathbb{E}_{np} (e^k - \mathbb{E}\, e^k) \right| + O\left(\max_{i,j} |\delta_{ij}| \right) \sum_{\ell=0}^{k-1} \mathbb{E}_{np} |e|^\ell. \qquad \text{(A.12)}
\end{aligned}$$

For all $k \in \{1, \cdots, m\}$, the strong law of large numbers with Theorem 2.5.7 in [9] 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 (A.12) 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

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

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

C.5 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*

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

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

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

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

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

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 D.3 over $|\mathcal{A}| \le k$. In fact, the proof is almost the same, with the only difference that bound

(A.23) should be replaced with the bound derived in Lemma 9(c); that is,

$$\max_{|\mathcal{A}|\leq k} \left\| n^{-1/2}\mathbf{E}_{\mathcal{A}} \right\|_2 \lesssim 1 \vee \left(kn^{-1}\log p\right)^{1/2}, \tag{A.13}$$

which holds with probability at least $1 - O(p^{-\nu})$. Notice that $\left(kn^{-1}\log p\right)^{1/2} \leq \log^{1/2} p$. Therefore, even if we use (A.13) instead of (A.23) 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 $(a\text{–}i)$ holds with probability at least $1 - O(\pi_{np})$.

For $(a\text{–}ii)$, we see that

$$\begin{aligned}
\left\|\mathbf{U}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{U}(\boldsymbol{\theta}^0)]\right\|_{\max} &\leq \left\|n^{-1}\mathbf{X}'\mathbf{X} - \mathbb{E}[n^{-1}\mathbf{X}'\mathbf{X}]\right\|_{\max} \\
&+ \left\|n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0) - \mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)]\right\|_{\max} \\
&+ 2\left\|n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0) - \mathbb{E}[n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)]\right\|_{\max} =: W_1 + W_2 + 2W_3.
\end{aligned} \tag{A.14}$$

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

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

Under Condition 3, we deduce

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

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

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

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

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

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

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

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

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

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

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

$$(b\text{–}ii) \quad \left\|\mathbf{v}(\boldsymbol{\theta}^0) - \mathbb{E}[\mathbf{v}(\boldsymbol{\theta}^0)]\right\|_{\max} \lesssim s(n^{-1}\log p)^{1/2}.$$

Similar to the proof of (a), we need to modify the proof of Lemma 8(b) in Section D.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

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

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

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

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

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

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

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

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.

C.6 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 (19) 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 $\boldsymbol{\delta}(:=\boldsymbol{\delta}(\boldsymbol{\theta})) := \hat{\boldsymbol{\beta}}^{\mathsf{aug}}(\boldsymbol{\theta}) - \boldsymbol{\beta}^{\mathsf{aug}}$. We start with introducing two inequalities

$$\sup_{\boldsymbol{\theta}\in\Theta_{np}} \left\| n^{-1}[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})]'\boldsymbol{\varepsilon} \right\|_{\max} \leq 2^{-1}\lambda, \tag{A.15}$$

$$\inf_{\boldsymbol{\theta}\in\Theta_{np}, \boldsymbol{\delta}\in\mathbb{V}} \boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \geq \sigma_e^2(1+o(1)), \tag{A.16}$$

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

$$\mathbb{V} = \left\{ \boldsymbol{\delta} \in \mathbb{R}^{2p} : \|\boldsymbol{\delta}_{\mathcal{S}^1}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{S}^0}\|_1,\ \|\boldsymbol{\delta}\|_0 \leq k \right\}. \tag{A.17}$$

It is well known that the rate of convergence of the Lasso estimator can be obtained provided that (A.15) and (A.16) hold. Thus we show that these two inequalities actually hold with high probability in Step 1, and then derive the convergence rate using (A.15) and (A.16) in Step 2.

Step 1. We check whether (A.15) and (A.16) actually hold with high probability. We first verify (A.15). By the proofs of Lemmas 8 and 4, we have

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

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 (A.24) in the proof of Lemma 8. That bound still holds with the same rate $O(n^{-1/2}\log p)$ even if we take $\mathcal{A} = \{1, \cdots, p\}$. Thus we conclude that (A.15) 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 (A.16), we derive the population lower bound first and then show that the

difference is negligible. From the construction, we have

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

$$\mathbb{E}[n^{-1}\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)'\mathbf{X}] = \mathbb{E}[n^{-1}\mathbf{X}'\widetilde{\mathbf{X}}(\boldsymbol{\theta}^0)] = \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime}.$$

Using these equations, we obtain the lower bound

$$\begin{aligned}
\inf_{\boldsymbol{\delta}\in\mathbb{V}} \boldsymbol{\delta}'\,\mathbb{E}\left[\mathbf{U}(\boldsymbol{\theta}^0)\right]\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 &= \inf_{\boldsymbol{\delta}\in\mathbb{V}} \boldsymbol{\delta}' \begin{pmatrix} \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} + \sigma_e^2\mathbf{I}_p & \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} \\ \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} & \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} + \sigma_e^2\mathbf{I}_p \end{pmatrix} \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \\
&= \inf_{\boldsymbol{\delta}\in\mathbb{V}} \boldsymbol{\delta}' \left\{ \begin{pmatrix} 1 & 1 \\ & \\ 1 & 1 \end{pmatrix} \otimes \boldsymbol{\Lambda}^0\boldsymbol{\Sigma}_f\boldsymbol{\Lambda}^{0\prime} + \sigma_e^2\mathbf{I}_{2p} \right\} \boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2^2 \\
&\geq \sigma_e^2. \qquad \text{(A.18)}
\end{aligned}$$

Because $\boldsymbol{\delta} \in \mathbb{V}$ is sparse and satisfies $|\mathcal{B}| \leq k$ for $\mathcal{B} := \mathrm{supp}(\boldsymbol{\delta})$, 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

$$\begin{aligned}
\sup_{|\mathcal{B}|\leq k,\,\boldsymbol{\theta}\in\Theta_{np}} \left\|\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}) - \mathbb{E}[\mathbf{U}_{\mathcal{B}}(\boldsymbol{\theta}^0)]\right\|_{\max} &= O\left(k^{1/2}\tilde{c}_{np}\right) \\
&= o(s^{-1}) \qquad \text{(A.19)}
\end{aligned}$$

with probability at least $1 - O(\pi_{np})$. Thus using (A.19), we have for any $\boldsymbol{\delta} \in \mathbb{V}$,

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

Rearranging the terms with (A.18) 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 (A.16). In consequence, two inequalities (A.15) and (A.16) hold with probability at least $1 - O(\pi_{np})$.

Step 2. This part is well known in the literature (e.g., [12]) 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

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

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 (A.15) yield

$$2^{-1}\boldsymbol{\delta}'\mathbf{U}(\boldsymbol{\theta})\boldsymbol{\delta} \leq \left\| n^{-1}\boldsymbol{\varepsilon}'[\mathbf{X}, \widetilde{\mathbf{X}}(\boldsymbol{\theta})] \right\|_{\max} \|\boldsymbol{\delta}\|_1 + \lambda\|\boldsymbol{\delta}\|_1 \leq (3/2)\lambda\|\boldsymbol{\delta}\|_1. \tag{A.20}$$

On the other hand, applying Lemma 1 of [12] to our model reveals that $\boldsymbol{\delta} \in \mathbb{V}$. Thus we can use (A.16), (A.20), and (A.17) 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.

D Additional technical lemmas and their proofs

D.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* $\mathbf{H} = (\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0/p)(\mathbf{F}^{0\prime}\hat{\mathbf{F}}/n)\mathbf{V}^{-1}$. *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 $\|\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0/p\|_2 \leq M$ and

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

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

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

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

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

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 [1]. 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., [11], Theorem 4.5.8) entails that all the r eigenvalues are positive for all n. Moreover, by Proposition 1 of [1] 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

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

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., [11], Theorem 4.3.1) yields

$$\begin{aligned}\lambda^r\left[(np)^{-1}\mathbf{X}\mathbf{X}'\right] &= \lambda^r\left[(np)^{-1}\left\{\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \left(\mathbf{E}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{E}'\right) + \mathbf{E}\mathbf{E}'\right\}\right] \\ &\geq \lambda^r\left[(np)^{-1}\mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{\Lambda}^0\mathbf{F}^{0\prime}\right] + \lambda^n\left[(np)^{-1}\left(\mathbf{E}\mathbf{\Lambda}^0\mathbf{F}^{0\prime} + \mathbf{F}^0\mathbf{\Lambda}^{0\prime}\mathbf{E}'\right)\right] + \lambda^n\left[(np)^{-1}\mathbf{E}\mathbf{E}'\right].\end{aligned} \tag{A.21}$$

The third term of lower bound (A.21) is obviously nonnegative. For the first term of lower bound (A.21), 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., [11], Theorem 4.2.6) yields

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

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 D.2, which together with the union bound establishes

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

for any $0\leq u\leq c$. 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 $n^{-1}\log p\leq c^{2}/(C\nu)$ without loss of generality. Therefore, the first term of lower bound (A.21) is bounded away from zero eventually. For the second term of (A.21), since the spectral norm gives the upper bound of the spectral radius we have

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

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

D.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 $\ell,i\in\{1,\cdots,n\}$ and $0\leq u\leq c$, we have

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

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

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

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

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

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

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

Proof. (a) To obtain the first result, we rely on the Hanson–Wright inequality. Let $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_m)' \in \mathbb{R}^m$ 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 $\mathbf{A} \in \mathbb{R}^{m \times m}$,

$$\mathbb{P}\left(\left|\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi} - \mathbb{E}\,\boldsymbol{\xi}'\mathbf{A}\boldsymbol{\xi}\right| > u\right) \leq 2\exp\left\{-\widetilde{C}_H \min\left(\frac{u^2}{K^4\|\mathbf{A}\|_F^2}, \frac{u}{K^2\|\mathbf{A}\|_2}\right)\right\}, \tag{A.22}$$

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 [13]). Using this inequality, we first prove the case when $\ell = i$. If we set $m = p$ and $\mathbf{A} = \text{diag}(p^{-1}, \cdots, p^{-1})$, 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 (A.22) 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 $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_{p+1})$, where $\mathbf{a}_1 = (0, p^{-1}, \cdots, p^{-1})'$ and $\mathbf{a}_j = \mathbf{0}$ for $j = 2, \cdots, p+1$. 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

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

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 (A.22) 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 [6]). Fix $k = 1$ and $j = 1$. Define $\mathcal{F}_{i-1}$ as the σ-field generated from $\{f_{\ell 1}^0 : \ell = i, i-1, \cdots\}$. 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 [13]), we have $V_i := \mathbb{E}\left[f_{ik}^{0\,2}e_{ij}^2 \mid \mathcal{F}_{i-1}\right] \leq 4C_e^2M^2$, 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 $i \in \{1, \cdots, n\}$,

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

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 [6] 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 $0 \leq u \leq MC_e^2$, 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 $0 \leq u \leq MC_e^2$ we have

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

(c) We prove the third inequality. Note that

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

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 $u \geq 0$ putting $C_J = M^2C_e^2$ leads to

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

(d) We show the last inequality. Note that for each k, $(f_{ik})_i \sim$ i.i.d. subG(M^2) since $|f_{ik}^0| \leq M$ a.s. by Lemma 1.8 of [13] under Condition 2. Thus the remaining is the same as (a). Set $C_K = 81M^4/\widetilde{C}_H$ here. Then for any $0 \leq u \leq 9M^2$, we have

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

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

D.3 Lemma 8 and its proof

Lemma 8 *Assume that Conditions 1–4 hold. Then for any set $\mathcal{A}$ satisfying $|\mathcal{A}| \leq k$, 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 $\|k^{-1/2}\mathbf{\Lambda}^0_{\mathcal{A}}\|_2 = O(1)$ for any $\mathcal{A}$ such that $|\mathcal{A}| \leq k$ under Condition 3, we first have

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

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

$$\left\| n^{-1/2} \mathbf{E}_{\boldsymbol{\eta}^0 \mathcal{A}} \right\|_2 \lesssim 1 \tag{A.23}$$

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

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

for any $u \geq 0$. 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 B.5 in the proofs below.

To prove (a), we have

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

Observe that U_1 is further bounded as

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

By Lemma 10, it is easy to see that

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

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

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

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

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

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

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

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

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

Thus we have

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

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

$$\begin{aligned}V_2 &\le \left\|\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta})-\widetilde{\mathbf{X}}_{\mathcal{A}}(\boldsymbol{\theta}^0)\right\|_{\max}\left\|n^{1/2}\boldsymbol{\varepsilon}\right\|_2\\ &\lesssim \left\|\mathbf{C}-\mathbf{C}^0\right\|_{\max}+\left\|\mathbf{E}_{\boldsymbol{\eta}}-\mathbf{E}_{\boldsymbol{\eta}^0}\right\|_{\max}\lesssim\tilde{c}_{np},\end{aligned}\tag{A.24}$$

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

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

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

B.4 Lemma 9 and its proof

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

(a) We have

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

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

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

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

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

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

Proof. Result (a) is obtained by Theorem 5.39 of [14]. Moreover, by the same theorem there exist some positive constants c and C such that for any $\mathcal{A}$ with $|\mathcal{A}| \leq k \leq n$ and every $t \geq 0$,

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

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

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

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.

B.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.

E Additional numerical details and results

E.1 Estimation procedure

In implementing the IPAD algorithm suggested in Section 2, we use the PC_{p1} criterion proposed in [2] 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 (8) 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 [3] and [4], respectively. Thus it is seen that neither BC-Knockoff 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. The value of the regularization parameter λ is chosen by the tenfold cross-validation. Then the LCD discussed in Section 2.2 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 [7]. 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 $\widehat{\text{MDA}}_j$, 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, $\widehat{\text{MDA}}_j$ 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 $|\widehat{\text{MDA}}_j| - |\widehat{\text{MDA}}_{j+p}|$, $j = 1, \cdots, p$.

E.2 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 1 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 1: 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

E.3 Methods of comparison in empirical analysis

We compare the following different methods in the empirical analysis presented in Section 5, 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 [2]. 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

$$y_t = \alpha_0 + \rho y_{t-1} + \boldsymbol{\delta}' \mathbf{z}_{t-1,\widehat{\mathcal{S}}} + \varepsilon_t, \tag{A.25}$$

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 (A.25) and use the mean value as the

final predicted value of y_{T+1}.

References

[1] Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–171.

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

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

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

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

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

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

[8] 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.

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

[10] Fan, Y., E. Demirkaya, G. Li, and J. Lv (2019). RANK: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association, to appear*.

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

[12] 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.

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

[14] 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.