SUPPLEMENTARY MATERIAL TO "INTERACTION PURSUIT IN HIGH-DIMENSIONAL MULTI-RESPONSE REGRESSION VIA DISTANCE CORRELATION"

By Yinfei Kong[¶], Daoji Li[‖], Yingying Fan[**] and Jinchi Lv[**]

California State University at Fullerton[¶], University of Central Florida[‖] and University of Southern California[**]

This Supplementary Material contains some intermediate steps of the proof of Theorem 1 and additional numerical studies and technical details, as well as the details about the post-screening interaction selection.

APPENDIX B: POST-SCREENING INTERACTION SELECTION

The screening step of IPDC can reduce the problem of interaction identification from a huge scale to a moderate one as shown in Section 2. In particular, the reduced interaction model can be of dimensionality smaller than the sample size. After the screening step, IPDC further selects important interactions and main effects. Thanks to the much reduced scale, the selection step can be conducted in a computationally efficient fashion by exploiting regularization methods for the multi-response regression. Various regularization methods have been developed for multi-response linear models. See, for example, [3], [7], [5], [6], [25], and references therein. Those methods were usually investigated for the scenario of no interactions. For the selection step of IPDC, we aim at interaction model recovery by employing a two-step variable selection procedure, where we first recover the support union using the idea of group variable selection [35] and then estimate the individual supports for each column of the regression coefficient matrix via an additional refitting step of Lasso [32] applied to the recovered support union (see, e.g., [14] for connections and differences among regularization methods).

To simplify the presentation, hereafter we assume that the response vector $\mathbf{y}$ is centered with mean zero and all interactions $X_k X_\ell$ are also centered to have mean zero with a slight abuse of notation, which eliminates the intercept vector $\boldsymbol{\alpha}$ in model (1). Thus given an i.i.d. sample $(\mathbf{y}_i, \mathbf{x}_i)_{i=1}^n$, the multi-response interaction model (1) can be rewritten in the matrix form

$$\mathbf{Y} = \widetilde{\mathbf{X}}\mathbf{B} + \mathbf{W}, \tag{A.1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)^T \in \mathbb{R}^{n\times q}$ is the response matrix, $\widetilde{\mathbf{X}} = (\mathbf{X}, \mathbf{Z}) \in \mathbb{R}^{n\times \widetilde{p}}$ with $\widetilde{p} = p(p+1)/2$ is the full augmented design matrix with $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$ the main effect matrix and $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T \in \mathbb{R}^{n\times [p(p-1)/2]}$ the interaction matrix, $\mathbf{z}_i$'s are defined similarly to $\mathbf{z}$, $\mathbf{B} = (\mathbf{B}_{\mathbf{x}}^T, \mathbf{B}_{\mathbf{z}}^T)^T \in \mathbb{R}^{\widetilde{p}\times q}$ is the regression coefficient matrix with $\mathbf{B}_{\mathbf{x}} \in \mathbb{R}^{p\times q}$ and $\mathbf{B}_{\mathbf{z}} \in \mathbb{R}^{[p(p-1)/2]\times q}$, and $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)^T \in \mathbb{R}^{n\times q}$ is the error matrix.

B.1. Interaction and main effect selection. Let S be the row support of the true regression coefficient matrix $\mathbf{B}^*$ in model (A.1), which corresponds to the index set of nonzero rows of $\mathbf{B}^*$; that is, if $k \in S$, then the kth row of $\mathbf{B}^*$ has at least one nonzero component. Denote by $\mathcal{J} = \{j_1, \cdots, j_{d_1}\} \subset \{1, \cdots, p\}$ and $\mathcal{K} = \{k_1, \cdots, k_{d_2}\} \subset \{1, \cdots, p\}$ the index sets of retained main effects and interaction variables after the screening step of IPDC, respectively. Then the reduced design matrix is $(\widetilde{\mathbf{x}}_{j_1}, \cdots, \widetilde{\mathbf{x}}_{j_{d_1}}, \widetilde{\mathbf{x}}_{k_1} \circ \widetilde{\mathbf{x}}_{k_2}, \cdots, \widetilde{\mathbf{x}}_{k_{d_2-1}} \circ \widetilde{\mathbf{x}}_{k_{d_2}}) \in \mathbb{R}^{n\times d}$ with $d = d_1 + d_2(d_2-1)/2$, where $\widetilde{\mathbf{x}}_\ell$ is the ℓth column of $\mathbf{X}$. Let $\widetilde{S} \subset \{1, \cdots, \widetilde{p}\}$ be the index set given by the columns of such a reduced matrix in the full matrix $\widetilde{\mathbf{X}}$. As guaranteed by Theorem 1, the true row support S can be contained in the reduced set $\widetilde{S}$ by IPDC with high probability that converges to one at a fast rate as sample size n increases.

Observe that the true row support S is the union of individual supports of the columns of the true regression coefficient matrix $\mathbf{B}^*$ corresponding to the q responses. Given any set $J \subset \{1, \cdots, \widetilde{p}\}$, denote by $\mathbf{B}_J$ a submatrix of $\mathbf{B}$ formed by the rows indexed by J. For the support union recovery, we exploit the multivariate group Lasso given by the following regularization problem

$$\min_{\mathbf{B}_{\widetilde{S}^c}=\mathbf{0}} \left\{ \frac{1}{2nq} \|\mathbf{Y} - \widetilde{\mathbf{X}}\mathbf{B}\|_F^2 + \lambda \|\mathbf{B}\|_{2,1} \right\}, \tag{A.2}$$

where $\widetilde{S}^c$ is the complement of the set $\widetilde{S}$, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\lambda \geq 0$ is a regularization parameter, and $\|\cdot\|_{2,1}$ stands for the matrix rowwise $(2,1)$-norm defined as $\|\mathbf{M}\|_{2,1} = \sum_i (\sum_j m_{ij}^2)^{1/2}$ for any matrix $\mathbf{M} = (m_{ij})$. Note that $\widetilde{S}$ and $\widehat{\mathcal{M}} \cup \widehat{\mathcal{I}}$ share the same cardinality. We should remark that as ensured by Theorem 2, the computational cost of solving the optimization problem (A.2) can be substantially reduced compared to that of solving the same optimization problem without the screening step, that is, with $\widetilde{S} = \{1, \cdots, \widetilde{p}\}$.

The multivariate group Lasso has been widely used in the multi-response linear regression models typically without interaction terms. For example,

[24] and [25] established the oracle inequalities for the case when the design matrix is deterministic and the error matrix has i.i.d. Gaussian entries. [28] investigated the model selection consistency in terms of support union recovery of the multivariate group Lasso under the assumptions that the design matrix is drawn with i.i.d. Gaussian row vectors and all the entries of the error matrix are i.i.d. Gaussian. We will relax such Gaussianity assumptions and justify that this group variable selection procedure continues to perform well in the presence of interactions.

Once the row support of the true regression coefficient matrix is recovered, it is straightforward to recover the individual supports of each column of the regression coefficient matrix by an additional refitting step of applying the ordinary Lasso to the recovered support union. Since the sampling properties of Lasso have been extensively studied and are now well understood in the literature, we will provide only theoretical analysis of the multivariate group Lasso problem (A.2).

B.2. Support union recovery and oracle inequalities. To facilitate our technical analysis for the selection step of IPDC, we impose a few additional regularity conditions.

CONDITION 4. *The covariate vector* $\mathbf{x}$ *has a sub-Gaussian distribution and* $s = |S| = O(n^{\xi})$ *for some constant* $0 \le \xi < 1/4$.

CONDITION 5 (RE(s) assumption). *There exists some positive constant* κ *such that*

$$\kappa(s) = \min_{|J|\le s,\, \mathbf{\Delta}\in\mathbb{R}^{\widetilde{p}\times q}\setminus\{\mathbf{0}\},\, \|\mathbf{\Delta}_{J^c}\|_{2,1}\le 3\|\mathbf{\Delta}_J\|_{2,1}} \frac{\|\mathbf{\Sigma}^{1/2}\mathbf{\Delta}\|_F}{\|\mathbf{\Delta}_J\|_F} \ge \kappa,$$

where $\mathbf{\Sigma}$ *is the covariance matrix of* $\widetilde{\mathbf{x}} = (\mathbf{x}^T, \mathbf{z}^T)^T$.

CONDITION 6. *The error vector* $\mathbf{w}$ *has a sub-exponential distribution.*

The first part of Condition 4 is a mild assumption on the distribution of the covariates. It can be satisfied by many light-tailed distributions such as Gaussian distributions and distributions with bounded support. The second part of Condition 4 puts a row sparsity constraint on the true regression coefficient matrix. In particular, the requirement of $\xi < 1/4$ reflects the difficulty of interaction selection in high dimensions.

Condition 5 is a natural extension of the restricted eigenvalue (RE) assumption introduced in [1] since here we use the rowwise $(2,1)$-norm in

place of the L_1-norm. The RE assumption has been commonly used to establish the oracle inequalities for the Lasso and Dantzig selector [4]. For simplicity we still refer to Condition 5 as the RE(s) assumption. This condition is also similar to Condition 3.1 in [24] and Condition 4.1 in [25], who considered the scenario of deterministic design matrix and no interactions. Condition 6 assumes the sub-exponential distribution for the error vector, which is key to establishing the deviation probability bound for $\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty}$. Here $\|\cdot\|_{2,\infty}$ denotes the matrix rowwise $(2,\infty)$-norm defined as $\|\mathbf{M}\|_{2,\infty} = \max_i(\sum_j m_{ij}^2)^{1/2}$ for any matrix $\mathbf{M} = (m_{ij})$. Hereafter p involved in the regularization parameter λ and probability bounds is understood implicitly as $\max\{n, p\}$.

THEOREM 3. *Assume that all the conditions of Theorem 1 and Conditions 4–6 hold, $q \le p$, $\log p = o(n^\eta)$ with $\eta = \min\{\eta_0, 1/2 - 2\xi\}$, and set $\lambda = c_3\sqrt{(\log p)/(nq)}$ with $c_3 > 0$ some constant. Then with probability at least $1 - O\{\exp(-Cn^{\eta_0/2})\} - O(p^{-c_4})$ for some constants $C, c_4 > 0$, the minimizer $\widehat{\mathbf{B}}$ of (A.2) satisfies*

$$(nq)^{-1/2}\|\widetilde{\mathbf{X}}(\widehat{\mathbf{B}} - \mathbf{B}^*)\|_F \le \frac{8c_3}{\kappa}\sqrt{s(\log p)/n}, \tag{A.3}$$

$$\frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \le \frac{64c_3}{\kappa^2}s\sqrt{(\log p)/n}. \tag{A.4}$$

If in addition $\min_{j\in S}\|\mathbf{B}_j^*\|/\sqrt{q} > 128c_3\kappa^{-2}s\sqrt{(\log p)/n}$, then with the same probability the row support of $\widetilde{\mathbf{B}}$ is identical to S, where the matrix $\widetilde{\mathbf{B}}$ is obtained by thresholding the jth row of $\widehat{\mathbf{B}}$ to zero for each j if $\|\widehat{\mathbf{B}}_j\|/\sqrt{q} \le 64c_3\kappa^{-2}s\sqrt{(\log p)/n}$. Moreover, if the RE($s$) assumption in Condition 5 is replaced by RE($2s$), then it holds with the same probability that

$$\frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F \le \frac{16\sqrt{10}c_3}{\kappa^2(2s)}\sqrt{s(\log p)/n}. \tag{A.5}$$

Theorem 3 establishes the model selection consistency of the IPDC followed by hard thresholding in terms of support union recovery. It also extends the oracle inequalities in Theorem 3.3 of [24] and Corollary 4.1 of [25] in three important aspects: the inclusion of interaction terms, the analysis of large random design matrix, and the relaxed distributional assumption. Such extensions make the technical analyses more involved and challenging. We should remark that the same results as in Theorem 3 hold with probability at least $1 - O(p^{-c_4})$ for the regularized estimator with $d = \widetilde{p}$, that is, without the screening step. It is also worth mentioning that the value

TABLE 5
Proportions of important main effects, important interaction, and all of them retained by different screening methods.

Method	X_{12}	X_{22}	X_1X_2	All
DCSIS2	1.00	1.00	0.06	0.06
DCSIS-square	0.00	0.81	1.00	0.00
IPDC	1.00	1.00	1.00	1.00

of the regularization parameter λ in our Theorem 3 is slightly larger than those used in [24] and [25], due to the more general model setting considered in our paper. In fact, such larger value of λ is needed to suppress the additional noise caused by the presence of interactions and the heavier-tailed distribution of model errors.

APPENDIX C: ADDITIONAL NUMERICAL STUDIES

C.1. Comparison of IPDC with individual components. Recall that the new interaction screening approach of IPDC treats the screening for interactions and the screening for main effects as two separate components. Since the distance correlation can capture nonlinear dependency between variables, a natural question is whether either of these two components might suffice for the purpose of the joint screening for both interactions and main effects. To ease the presentation, the component for interaction screening is referred to as DCSIS-square, and the component for main effect screening is called DCSIS2 as described in Section 3.1. Thus it is of interest to compare IPDC with both DCSIS2 and DCSIS-square. To this end, we revisit the setting 3 of Model 4 investigated in Section 3.1; see Table 1 for the screening performance of DCSIS2 and IPDC.

Table 5 reports the comparison results of these three methods. We see that DCSIS2, which is designed specifically for main effect screening, fails to retain the important interaction X_1X_2, and DCSIS-square, which is designed specifically for interaction screening, fails to retain the important main effect X_{12}. In contrast, the IPDC combines the strengths of its two individual components in screening for both interactions and main effects. Such an observation is in line with a key message spelled out in the Introduction, that is, a separate screening step for interactions can significantly enhance the screening performance if one aims at finding important interactions.

C.2. Performance of interaction and main effect selection.

C.2.1. *Selection in single-response models.* After the screening step, we further investigate the performance of selection for the interactions and main effects in the reduced feature space. The selection step in the single-response examples (Models 1–4) is implemented by the Lasso. Thus we refer to each two-stage interaction screening and selection procedure by the SIS2-Lasso, DCSIS2-Lasso, SIRI-Lasso, IP-Lasso, and IPDC-Lasso, respectively. The oracle procedure, which assumes that the true underlying sparse interaction model is known in advance, is used as a benchmark for comparison. In particular, in Model 4 the indicator covariate $\mathbb{I}(X_{12} \geq 0)$ instead of the linear predictor X_{12} is used in the oracle procedure.

Three performance measures, the prediction error (PE), the number of false positives (FP), and the number of false negatives (FN), are employed to assess the variable selection performance of each method in the single-response examples. The PE is defined as $E(Y - \widehat{Y})^2$ with $\widehat{Y}$ the predicted response. We generate an independent test sample of size $10,000$ to calculate the PE. The FP is defined as the total number of unimportant interactions and main effects included in the final model, while the FN is defined as the total number of important interactions and main effects missed by the final model.

Table 6 presents the means and standard errors of different performance measures. Since the screening results by all methods in Model 1 under three different settings are almost identical, one can expect that the corresponding selection results should be very similar, which is indeed the case. Thus we omit the selection results for Model 1 to save space. For Model 2, the performance of all methods is very similar across three settings. As a result, we only present the selection results under setting 2 of this model in Table 7 as a representative. The complete results are available upon request. Based on Tables 6 and 7, the following observations can be made.

- In Model 2, we see that IPDC-Lasso performs the best and closest to the oracle procedure across all measures, as shown in Table 7. For Model 3 under all settings, our method IPDC-Lasso has far lower mean prediction error than all other methods except for the oracle, according to Table 6. The advantage of IPDC-Lasso over other methods is also evident in Model 4.
- We remark that the gap between the prediction errors of the IPDC-Lasso and oracle in Model 4 is mainly because as mentioned before, the latter exploits the indicator covariate $\mathbb{I}(X_{12} \geq 0)$ whereas such prior information is unavailable to all other procedures. Even in this scenario of model misspecification, our method still performs well in identifying important interactions and main effects.

TABLE 6
Means and standard errors (in parentheses) of different selection performance measures for Models 3 and 4 over 100 replications.

Method	Model 3			Model 4		
	PE	FP	FN	PE	FP	FN
			Setting 1: $(p, \rho) = (2000, 0.5)$			
SIS2-Lasso	33.79 (0.87)	39.25 (3.79)	1.88 (0.04)	15.00 (0.32)	22.05 (2.50)	1.23 (0.06)
DCSIS2-Lasso	3.94 (0.41)	0.51 (0.17)	0.32 (0.05)	3.36 (0.38)	4.16 (0.72)	0.11 (0.04)
SIRI-Lasso	3.54 (0.40)	0.54 (0.36)	0.27 (0.05)	4.34 (0.49)	1.63 (0.40)	0.31 (0.07)
IP-Lasso	2.08 (0.28)	0.46 (0.11)	0.10 (0.03)	2.38 (0.05)	4.24 (0.64)	0.07 (0.03)
IPDC-Lasso	1.27 (0.10)	0.63 (0.20)	0.01 (0.01)	2.27 (0.02)	3.32 (0.47)	0.01 (0.01)
Oracle	1.017 (0.002)	0 (0)	0 (0)	1.022 (0.002)	0 (0)	0 (0)
			Setting 2: $(p, \rho) = (5000, 0.5)$			
SIS2-Lasso	36.31 (0.51)	61.78 (2.85)	1.97 (0.02)	15.27 (0.24)	39.29 (2.33)	1.15 (0.04)
DCSIS2-Lasso	5.81 (0.44)	1.20 (0.55)	0.54 (0.05)	4.17 (0.48)	3.45 (0.53)	0.20 (0.05)
SIRI-Lasso	4.48 (0.45)	0.42 (0.16)	0.37 (0.05)	4.70 (0.52)	2.24 (0.44)	0.37 (0.07)
IP-Lasso	2.52 (0.33)	1.83 (0.61)	0.15 (0.04)	2.51 (0.06)	6.75 (1.24)	0.14 (0.04)
IPDC-Lasso	1.38 (0.16)	0.91 (0.31)	0.02 (0.01)	2.30 (0.02)	4.39 (0.63)	0.01 (0.01)
Oracle	1.009 (0.002)	0 (0)	0 (0)	1.014 (0.002)	0 (0)	0 (0)
			Setting 3: $(p, \rho) = (2000, 0.1)$			
SIS2-Lasso	21.96 (0.19)	22.84 (3.18)	1.98 (0.01)	13.15 (0.10)	15.68 (2.04)	1.24 (0.05)
DCSIS2-Lasso	18.85 (0.47)	9.32 (2.47)	1.70 (0.05)	12.58 (0.28)	8.51 (1.73)	1.15 (0.05)
SIRI-Lasso	14.45 (0.72)	0.40 (0.17)	1.28 (0.07)	11.55 (0.45)	1.35 (0.43)	1.28 (0.07)
IP-Lasso	6.23 (0.63)	4.20 (1.37)	0.46 (0.06)	2.54 (0.17)	6.28 (1.54)	0.05 (0.02)
IPDC-Lasso	3.08 (0.44)	0.99 (0.21)	0.17 (0.04)	2.26 (0.01)	4.00 (0.79)	0.00 (0.00)
Oracle	1.017 (0.002)	0 (0)	0 (0)	1.022 (0.002)	0 (0)	0 (0)

TABLE 7
Means and standard errors (in parentheses) of different selection performance measures for setting 2 of Model 2 over 100 replications.

Method	PE	FP	FN
SIS2-Lasso	25.57 (1.61)	30.20 (3.31)	1.62 (0.10)
DCSIS2-Lasso	3.20 (0.40)	1.85 (0.44)	0.21 (0.04)
SIRI-Lasso	3.03 (0.38)	1.30 (0.23)	0.20 (0.04)
IP-Lasso	4.05 (0.45)	4.79 (1.06)	0.33 (0.05)
IPDC-Lasso	1.61 (0.20)	2.55 (0.49)	0.04 (0.02)
Oracle	1.014 (0.002)	0 (0)	0 (0)

TABLE 8
Means and standard errors (in parentheses) of different selection performance measures for Model 5 over 100 replications.

Method	PE	FP.main	FP.int	FN.main	FN.int
SIS.max-GLasso	6.24 (0.06)	127.64 (4.72)	699.08 (27.37)	1.22 (0.23)	9.50 (0.12)
SIS.sum-GLasso	6.12 (0.08)	185.28 (3.64)	810.72 (20.19)	0.74 (0.14)	9.06 (0.17)
DCSIS-GLasso	4.29 (0.14)	157.74 (3.98)	830.19 (23.00)	0.50 (0.10)	5.63 (0.29)
IPDC-GLasso	2.74 (0.09)	124.77 (3.02)	878.93 (20.60)	0.04 (0.02)	2.46 (0.23)
SIS.max-GLasso-Lasso	5.11 (0.05)	11.92 (0.84)	52.40 (3.17)	3.44 (0.19)	9.60 (0.10)
SIS.sum-GLasso-Lasso	4.99 (0.07)	15.75 (0.86)	63.73 (3.39)	3.07 (0.20)	9.35 (0.16)
DCSIS-GLasso-Lasso	3.40 (0.12)	11.87 (0.70)	62.98 (2.94)	1.41 (0.17)	6.36 (0.28)
IPDC-GLasso-Lasso	2.08 (0.09)	7.42 (0.36)	58.95 (2.32)	0.37 (0.09)	3.28 (0.24)
Oracle	1.048 (0.001)	0 (0)	0 (0)	0 (0)	0 (0)

C.2.2. *Selection in multi-response model.* As mentioned before, interaction and main effect selection in the multi-response model setting (Model 5) is conducted through a two-step procedure in the reduced feature space obtained by screening. Such a method first selects rows of the regression coefficient matrix using the group Lasso, and then applies the Lasso to each individual response for further selection of the rows. The goal of the individual Lasso is to eliminate the unimportant interactions and main effects that are included in the model recovered by the group Lasso. The resulting interaction screening and selection procedures are referred to as the SIS.max-GLasso-Lasso, SIS.sum-GLasso-Lasso, DCSIS-GLasso-Lasso, and IPDC-GLasso-Lasso, respectively. We also include for comparison the procedures that exploit only the group Lasso in the selection step, which are named by dropping the Lasso component. The oracle procedure is the ordinary least-squares estimation applied to each response separately on the corresponding true support.

The same performance measures as defined in Section C.2.1 are employed to evaluate different methods, except that the PE is now calculated as the average prediction error across all $q = 10$ responses. To further differentiate the false positives and false negatives for the main effects and interactions, we attach ".main" and ".int" to both measures of FP and FN as shown in Table 8.

Table 8 reports the selection results for Model 5. The FP.int is relatively large for all methods since even after screening, there are still a large number of interactions left, due to the presence of multiple responses. We observe that a further step of individual Lasso implemented on the support of group lasso for each response separately can substantially reduce the FP for both interactions and main effects. Moreover, our method IPDC-GLasso-Lasso outperforms all competitors under all performance measures.

TABLE 9
Means and standard errors (in parentheses) of prediction error as well as numbers of selected main effects and interactions for each method in mice data.

Method	PE ($\times 10^{-3}$)	Model Size	
		Main	Interaction
SIS2-Lasso	100.14 (5.57)	0.14 (0.03)	9.56 (0.11)
DCSIS2-Lasso	100.91 (6.12)	0.12 (0.03)	9.22 (0.11)
SIRI-Lasso	247.53 (10.11)	1.07 (0.03)	0.46 (0.13)
IP-Lasso	101.34 (4.99)	5.00 (0.09)	8.55 (0.21)
IPDC-Lasso	96.55 (5.40)	3.04 (0.07)	7.31 (0.11)

C.3. Univariate gene expression study. We study the the inbred mouse microarray gene expression data set in Lan et al. [21]. There are 60 mouse arrays, with 31 from female mice and 29 from male mice. The response variable is the gene expression level of stearoyl-CoA desaturase-1 (Scd1), a gene involved in fat storage. Specifically, Scd1 controls lipid metabolism and insulin sensitivity. The covariates are gene expression levels for 22,690 of the mice's other genes. Therefore, the sample size $n = 60$, the number of covariates $p = 22,690$, and the number of responses $q = 1$. All variables involved in this studey are continuous.

This data set is publicly available on the Gene Expression Omnibus website (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330), and has been studied in Hao and Zhang [16] and Narisetty and He [27]. Following Narisetty and He [27], we randomly split the data into training and test sets of sizes 55 and 5, respectively. Furthermore, to ameliorate the numerical instability caused by the relatively small sample size we perform 200 random splits and calculate the mean prediction errors and the corresponding standard errors to better evaluate the performance of various methods. We compare the IPDC with the SIS2, DCSIS2, SIRI, and IP. Detailed descriptions of all these methods can be found in Sections 3 and C.2.

The final selection results on the prediction error and selected model size are summarized in Table 9. We see that IPDC-Lasso performs noticeably better than its competitors. Further, paired t-tests of prediction errors on the 200 splits of IPDC-Lasso against SIS2-Lasso, DCSIS2-Lasso, SIRI-Lasso, and IP-Lasso lead to p-values 3.40×10^{-2}, 1.94×10^{-2}, 1.90×10^{-36}, and 2.64×10^{-2}, respectively. These test results demonstrate the significantly improved performance of IPDC over other methods at the 5% level.

APPENDIX D: ADDITIONAL PROOFS OF MAIN RESULTS

D.1. Step 1.3 of Part 1 in the proof of Theorem 1. We now consider the term $\widehat{T}^*_{k1,3} - T_{k1,3}$. Applying the Cauchy-Schwarz inequality twice leads to

$$\begin{aligned} T_{k1,3} &\le \left\{E\left[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)\right] E\left[\phi^2(X_{1k}^*, X_{2k}^*)\mathbb{I}\{\phi(X_{1k}^*, X_{2k}^*) > M_1\}\right]\right\}^{1/2} \\ &\le E^{1/2}[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)] \left\{E[\phi^4(X_{1k}^*, X_{2k}^*)] P\{\phi(X_{1k}^*, X_{2k}^*) > M_1\}\right\}^{1/4}. \end{aligned} \tag{A.6}$$

In view of (17), we have

$$\begin{aligned} E[\phi^4(X_{1k}^*, X_{2k}^*)] &\le E[(X_{1k}^2 + X_{2k}^2)^4] \le E\{[2(X_{1k}^4 + X_{2k}^4)]^2\} \\ &\le E[8(X_{1k}^8 + X_{2k}^8)] = 16E(X_{1k}^8) \end{aligned} \tag{A.7}$$

and by Bonferroni's inequality,

$$\begin{aligned} P\{\phi(X_{1k}^*, X_{2k}^*) > M_1\} &\le P(X_{1k}^2 + X_{2k}^2 > M_1) \le P(X_{1k}^2 > M_1/2) \\ &\quad + P(X_{2k}^2 > M_1/2) = 2P(X_{1k}^2 > M_1/2) \\ &\le 2\exp(-c_0 M_1/2) E[\exp(c_0 X_{1k}^2)]. \end{aligned} \tag{A.8}$$

Combining (25) with (A.6)–(A.8) and by Condition 2, we obtain $T_{k1,3} \le \widetilde{C}_3 \exp(-8^{-1}c_0 M_1)$, where $\widetilde{C}_3$ is some positive constant. Since $M_1 = n^{\xi_1}$, it holds that for any positive constant $\widetilde{C}$,

$$0 \le T_{k1,3} \le \widetilde{C}_3 \exp(-8^{-1}c_0 n^{\xi_1}) \le \widetilde{C} n^{-\kappa_2}/48 \tag{A.9}$$

for all $1 \le k \le p$ when n is sufficiently large. This entails that

$$\begin{aligned} &P(\max_{1\le k\le p} |\widehat{T}^*_{k1,3} - T_{k1,3}| \ge \widetilde{C} n^{-\kappa_2}/24) \\ &\quad \le P(\max_{1\le k\le p} |\widehat{T}^*_{k1,3}| \ge \widetilde{C} n^{-\kappa_2}/48) \end{aligned} \tag{A.10}$$

for all n sufficiently large. Thus applying Markov's inequality and noting that $\widehat{T}^*_{k1,3} \ge 0$ and $E(\widehat{T}^*_{k1,3}) = T_{k1,3}$, we have

$$\begin{aligned} P(|\widehat{T}^*_{k1,3}| \ge \delta/2) &\le (\delta/2)^{-1} E(|\widehat{T}^*_{k1,3}|) \le (\delta/2)^{-1} E(\widehat{T}^*_{k1,3}) \\ &= (\delta/2)^{-1} T_{k1,3} \end{aligned}$$

for any $\delta > 0$. Choosing $\delta = \widetilde{C} n^{-\kappa_2}/24$ in the above inequality and in view of (A.9), it follows that $P(|\widehat{T}^*_{k1,3}| \ge \widetilde{C} n^{-\kappa_2}/48) \le 48\widetilde{C}^{-1}\widetilde{C}_3 n^{\kappa_2} \exp(-8^{-1}c_0 n^{\xi_1})$. This inequality together with (A.10) and Bonferroni's inequality yields

$$\begin{aligned} &P(\max_{1\le k\le p} |\widehat{T}^*_{k1,3} - T_{k1,3}| \ge \widetilde{C} n^{-\kappa_2}/24) \\ &\quad \le 48p\widetilde{C}^{-1}\widetilde{C}_3 n^{\kappa_2} \exp(-8^{-1}c_0 n^{\xi_1}). \end{aligned} \tag{A.11}$$

D.2. Step 2 of Part 1 in the proof of Theorem 1. In this step, we handle the term $\max_{1\le k\le p}|\widehat{T}_{k2}-T_{k2}|$. Define $T_{k2,1}=E\left[\phi(X_{1k}^*,X_{2k}^*)\right]$ and $T_{k2,2}=E\left[\psi(\mathbf{y}_1^*,\mathbf{y}_2^*)\right]$. Then $T_{k2}=T_{k2,1}T_{k2,2}$. Similarly, $\widehat{T}_{k2}$ can be rewritten as $\widehat{T}_{k2}=\widehat{T}_{k2,1}\widehat{T}_{k2,2}$ by letting $\widehat{T}_{k2,1}=n^{-2}\sum_{i,j=1}^n\phi(X_{ik}^*,X_{jk}^*)$ and $\widehat{T}_{k2,2}=n^{-2}\sum_{i,j=1}^n\psi(\mathbf{y}_i^*,\mathbf{y}_j^*)$. An application of similar arguments as in Step 1 results in that for any positive constant $\widetilde{C}$, there exist some positive constants $\widetilde{C}_1,\cdots,\widetilde{C}_4$ such that

$$P\big(|\widehat{T}_{k2,1}-T_{k2,1}|\ge\widetilde{C}n^{-\kappa_2}/4\big)\le\widetilde{C}_1\exp\{-\widetilde{C}_2n^{(1-2\kappa_2)/3}\},$$
$$P\big(|\widehat{T}_{k2,2}-T_{k2,2}|\ge\widetilde{C}n^{-\kappa_2}/4\big)\le\widetilde{C}_3\exp\{-\widetilde{C}_4n^{(1-2\kappa_2)/5}\}.$$

In view of (17) and (18), we have $T_{k2,1}\le 2E(X_{1k}^2)$ and $T_{k2,2}\le 2E(\|\widetilde{\mathbf{y}}_1\|^2)$. By Condition 2, $T_{k2,1}$ and $T_{k2,2}$ are uniformly bounded from above by some positive constant for all $1\le k\le p$. Thus it follows from Lemma 1 that for any positive constant $\widetilde{C}$, there exist some positive constants $\widetilde{C}_5$ and $\widetilde{C}_6$ such that

$$\begin{aligned}P\big(|\widehat{T}_{k2}-T_{k2}|\ge\widetilde{C}n^{-\kappa_2}/4\big)&=P\big(|\widehat{T}_{k2,1}\widehat{T}_{k2,2}-T_{k2,1}T_{k2,2}|\ge\widetilde{C}n^{-\kappa_2}/4\big)\\&\le\widetilde{C}_5\exp\{-\widetilde{C}_6n^{(1-2\kappa_2)/5}\}\end{aligned}$$

for all $1\le k\le p$. By Bonferroni's inequality, we obtain

$$\begin{aligned}P\Big(\max_{1\le k\le p}|\widehat{T}_{k2}-T_{k2}|\ge\widetilde{C}n^{-\kappa_2}/4\Big)&\le\sum_{k=1}^pP\big(|\widehat{T}_{k2}-T_{k2}|\ge\widetilde{C}n^{-\kappa_2}/4\big)\\&\le p\widetilde{C}_5\exp\{-\widetilde{C}_6n^{(1-2\kappa_2)/5}\}.\end{aligned}\tag{A.12}$$

D.3. Step 3 of Part 1 in the proof of Theorem 1. We now consider the term $\widehat{T}_{k3}-T_{k3}$. Define a U-statistic

$$\widehat{T}_{k3}^*=6[n(n-1)(n-2)]^{-1}\sum_{i<j<l}g(X_{ik}^*,\mathbf{y}_i^*;X_{jk}^*,\mathbf{y}_j^*;X_{lk}^*,\mathbf{y}_l^*)$$

with the kernel $g(X_{ik}^*,\mathbf{y}_i^*;X_{jk}^*,\mathbf{y}_j^*;X_{lk}^*,\mathbf{y}_l^*)$ given by

$$\begin{aligned}g(X_{ik}^*,\mathbf{y}_i^*;X_{jk}^*,\mathbf{y}_j^*;X_{lk}^*,\mathbf{y}_l^*)&=\phi(X_{ik}^*,X_{jk}^*)\psi(\mathbf{y}_i^*,\mathbf{y}_l^*)+\phi(X_{ik}^*,X_{lk}^*)\psi(\mathbf{y}_i^*,\mathbf{y}_j^*)\\&+\phi(X_{jk}^*,X_{ik}^*)\psi(\mathbf{y}_j^*,\mathbf{y}_l^*)+\phi(X_{jk}^*,X_{lk}^*)\psi(\mathbf{y}_j^*,\mathbf{y}_i^*)\\&+\phi(X_{lk}^*,X_{ik}^*)\psi(\mathbf{y}_l^*,\mathbf{y}_j^*)+\phi(X_{lk}^*,X_{jk}^*)\psi(\mathbf{y}_l^*,\mathbf{y}_i^*).\end{aligned}$$

Then $\widehat{T}_{k3} = n^{-2}(n-1)(n-2)[\widehat{T}^*_{k3} + (n-2)^{-1}\widehat{T}^*_{k1}]$. By the triangle inequality, we deduce

$$
\begin{aligned}
|\widehat{T}_{k3} - T_{k3}| = \Big| & \frac{(n-1)(n-2)}{n^2}(\widehat{T}^*_{k3} - T_{k3}) - \frac{3n-2}{n^2}T_{k3} \\
& + \frac{n-1}{n^2}(\widehat{T}^*_{k1} - T_{k1}) + \frac{n-1}{n^2}T_{k1}\Big| \le |\widehat{T}^*_{k3} - T_{k3}| \\
& + |\frac{3}{n}T_{k3}| + |\widehat{T}^*_{k1} - T_{k1}| + |\frac{1}{n}T_{k1}|.
\end{aligned} \tag{A.13}
$$

It follows from the Cauchy-Schwarz inequality and (17)–(18) that

$$
\begin{aligned}
T_{k3} &\le \left\{E[\phi^2(X^*_{1k}, X^*_{2k})]E[\psi^2(\mathbf{y}^*_1, \mathbf{y}^*_3)]\right\}^{1/2} \\
&\le \left\{E\left[(X^2_{1k} + X^2_{2k})^2\right] E\left[(\|\widetilde{\mathbf{y}}_1\|^2 + \|\widetilde{\mathbf{y}}_3\|^2)^2\right]\right\}^{1/2} \\
&\le \left\{E[2(X^4_{1k} + X^4_{2k})]E[2(\|\widetilde{\mathbf{y}}_1\|^4 + \|\widetilde{\mathbf{y}}_3\|^4)]\right\}^{1/2} \\
&= 4\left\{E(X^4_{1k})E(\|\widetilde{\mathbf{y}}_1\|^4)\right\}^{1/2}
\end{aligned}
$$

for all $1 \le k \le p$.

In Step 1, we have shown that $T_{k1} \le 4\left\{E(X^4_{1k})E(\|\widetilde{\mathbf{y}}_1\|^4)\right\}^{1/2}$ for all $1 \le k \le p$. By Condition 2, $E(X^4_{1k})$ and $E(\|\mathbf{y}_1\|^4)$ are uniformly bounded from above by some positive constant for all $1 \le k \le p$. Note that $T_{k1} \ge 0$ and $T_{k3} \ge 0$. Thus for any positive constant $\widetilde{C}$, we have $\max_{1\le k\le p}|3n^{-1}T_{k3}| < \widetilde{C}n^{-\kappa_2}/16$ and $\max_{1\le k\le p}|n^{-1}T_{k1}| < \widetilde{C}n^{-\kappa_2}/16$ for all n sufficiently large. These two inequalities along with (A.13) entail

$$
\begin{aligned}
P(\max_{1\le k\le p}|\widehat{T}_{k3} - T_{k3}| \ge \widetilde{C}n^{-\kappa_2}/4) &\le P(\max_{1\le k\le p}|\widehat{T}^*_{k3} - T_{k3}| \\
&\ge \widetilde{C}n^{-\kappa_2}/16) + P(\max_{1\le k\le p}|\widehat{T}^*_{k1} - T_{k1}| \ge \widetilde{C}n^{-\kappa_2}/16).
\end{aligned} \tag{A.14}
$$

Replacing $\widetilde{C}$ with $\widetilde{C}/2$ in (31) gives

$$
\begin{aligned}
P(\max_{1\le k\le p}|\widehat{T}^*_{k1} - T_{k1}| \ge \widetilde{C}n^{-\kappa_2}/16) &\le p\widetilde{C}_1\exp\{-\widetilde{C}_2 n^{(1-2\kappa_2)/3-2\eta}\} \\
&+ \widetilde{C}_3\exp\{-\widetilde{C}_4 n^{3\eta/2}\},
\end{aligned} \tag{A.15}
$$

where $\widetilde{C}_1, \cdots, \widetilde{C}_4$ are some positive constants.

It remains to bound $P(\max_{1\le k\le p}|\widehat{T}^*_{k3} - T_{k3}| \ge \widetilde{C}n^{-\kappa_2}/16)$. Let $T_{k3} = T_{k3,1} + T_{k3,2} + T_{k3,3}$ with

$$
\begin{aligned}
T_{k3,1} &= E\left[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_3)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_1, \mathbf{y}^*_3) \le M_4\}\right], \\
T_{k3,2} &= E\left[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_3)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_1, \mathbf{y}^*_3) > M_4\}\right], \\
T_{k3,3} &= E\left[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_3)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) > M_3\}\right].
\end{aligned}
$$

Similarly, write $\widehat{T}^*_{k3}$ as $\widehat{T}^*_{k3} = \widehat{T}^*_{k3,\,1} + \widehat{T}^*_{k3,\,2} + \widehat{T}^*_{k3,\,3}$, where

$$\begin{aligned}
\widehat{T}^*_{k3,\,1} =& \frac{1}{n(n-1)(n-2)} \sum_{i<j<l} \big[\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) \le M_3\} \\
& \quad \cdot \mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_l) \le M_4\} \\
& + \phi(X^*_{ik}, X^*_{lk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{lk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_j) \le M_4\} \\
& + \phi(X^*_{jk}, X^*_{ik})\psi(\mathbf{y}^*_j, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{jk}, X^*_{ik}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_j, \mathbf{y}^*_l) \le M_4\} \\
& + \phi(X^*_{jk}, X^*_{lk})\psi(\mathbf{y}^*_j, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{jk}, X^*_{lk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_j, \mathbf{y}^*_i) \le M_4\} \\
& + \phi(X^*_{lk}, X^*_{ik})\psi(\mathbf{y}^*_l, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{lk}, X^*_{ik}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_l, \mathbf{y}^*_j) \le M_4\} \\
& + \phi(X^*_{lk}, X^*_{jk})\psi(\mathbf{y}^*_l, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{lk}, X^*_{jk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_l, \mathbf{y}^*_i) \le M_4\}\big] \\
=:& \frac{6}{n(n-1)(n-2)} \sum_{i<j<l} \widetilde{g}(X^*_{ik}, \mathbf{y}^*_i; X^*_{jk}, \mathbf{y}^*_j; X^*_{lk}, \mathbf{y}^*_l),
\end{aligned}$$

$$\begin{aligned}
\widehat{T}^*_{k3,\,2} =& \frac{1}{n(n-1)(n-2)} \sum_{i<j<l} \big[\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) \le M_3\} \\
& \quad \cdot \mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_l) > M_4\} \\
& + \phi(X^*_{ik}, X^*_{lk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{lk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_j) > M_4\} \\
& + \phi(X^*_{jk}, X^*_{ik})\psi(\mathbf{y}^*_j, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{jk}, X^*_{ik}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_j, \mathbf{y}^*_l) > M_4\} \\
& + \phi(X^*_{jk}, X^*_{lk})\psi(\mathbf{y}^*_j, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{jk}, X^*_{lk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_j, \mathbf{y}^*_i) > M_4\} \\
& + \phi(X^*_{lk}, X^*_{ik})\psi(\mathbf{y}^*_l, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{lk}, X^*_{ik}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_l, \mathbf{y}^*_j) > M_4\} \\
& + \phi(X^*_{lk}, X^*_{jk})\psi(\mathbf{y}^*_l, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{lk}, X^*_{jk}) \le M_3\}\mathbb{I}\{\psi(\mathbf{y}^*_l, \mathbf{y}^*_i) > M_4\}\big],
\end{aligned}$$

$$\begin{aligned}
\widehat{T}^*_{k3,\,3} =& \frac{1}{n(n-1)(n-2)} \sum_{i<j<l} \big[\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) > M_3\} \\
& + \phi(X^*_{ik}, X^*_{lk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{lk}) > M_3\} \\
& + \phi(X^*_{jk}, X^*_{ik})\psi(\mathbf{y}^*_j, \mathbf{y}^*_l)\mathbb{I}\{\phi(X^*_{jk}, X^*_{ik}) > M_3\} \\
& + \phi(X^*_{jk}, X^*_{lk})\psi(\mathbf{y}^*_j, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{jk}, X^*_{lk}) > M_3\} \\
& + \phi(X^*_{lk}, X^*_{ik})\psi(\mathbf{y}^*_l, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{lk}, X^*_{ik}) > M_3\} \\
& + \phi(X^*_{lk}, X^*_{jk})\psi(\mathbf{y}^*_l, \mathbf{y}^*_i)\mathbb{I}\{\phi(X^*_{lk}, X^*_{jk}) > M_3\}\big].
\end{aligned}$$

Clearly, $\widehat{T}^*_{k3,\,1}$, $\widehat{T}^*_{k3,\,2}$, and $\widehat{T}^*_{k3,\,3}$ are unbiased estimators of $T_{k3,\,1}$, $T_{k3,\,2}$, and

$T_{k3,3}$, respectively. By the triangle inequality, we have

$$P(\max_{1\le k\le p}|\widehat{T}^*_{k3}-T_{k3}|\ge \widetilde{C}n^{-\kappa_2}/16) \le \sum_{j=1}^{3} P(\max_{1\le k\le p}|\widehat{T}^*_{k3,j}-T_{k3,j}|\ge \widetilde{C}n^{-\kappa_2}/48). \tag{A.16}$$

Note that $\widetilde{g}$ defined in the expression for $\widehat{T}^*_{k3,1}$ is the kernel of the U-statistic $\widehat{T}^*_{k3,1}$ of order 3. Applying similar arguments to those for dealing with $\widehat{T}^*_{k1,1}$ in Step 1 yields

$$\begin{aligned} P(|\widehat{T}^*_{k3,1}-T_{k3,1}|\ge \widetilde{C}n^{-\kappa_2}/48) &\le 2\exp\{-m_3\widetilde{C}^2n^{-2\kappa_2}/(1152M_3^2M_4^2)\} \\ &\le 2\exp\{-\widetilde{C}_5 n^{1-2\kappa_2-2\xi_3-2\xi_4}\} \end{aligned}$$

with $\widetilde{C}_5$ some positive constant, by setting $M_3=n^{\xi_3}$ and $M_4=n^{\xi_4}$ with $\xi_3,\xi_4>0$ and noting that $m_3=\lfloor n/3\rfloor$. Thus it follows from Bonferroni's inequality that

$$\begin{aligned} &P(\max_{1\le k\le p}|\widehat{T}^*_{k3,1}-T_{k3,1}|\ge \widetilde{C}n^{-\kappa_2}/48) \\ &\le \sum_{1\le k\le p} P(|\widehat{T}^*_{k3,1}-T_{k3,1}|\ge \widetilde{C}n^{-\kappa_2}/48) \\ &\le 2p\exp\{-\widetilde{C}_5 n^{1-2\kappa_2-2\xi_3-2\xi_4}\}. \end{aligned} \tag{A.17}$$

Using similar arguments to those for (29)–(30), we can show that

$$P(\max_{1\le k\le p}|\widehat{T}^*_{k3,2}-T_{k3,2}|\ge \widetilde{C}n^{-\kappa_2}/48) \le \widetilde{C}_6 n^{\kappa_2+\xi_3}\exp\{-2^{-3/2}c_0n^{\xi_4/2}\}, \tag{A.18}$$

$$P(\max_{1\le k\le p}|\widehat{T}^*_{k3,3}-T_{k3,3}|\ge \widetilde{C}n^{-\kappa_2}/48) \le p\widetilde{C}_7 n^{\kappa_2}\exp(-8^{-1}c_0n^{\xi_3}), \tag{A.19}$$

where $\widetilde{C}_6$ and $\widetilde{C}_7$ are some positive constants.

Combining the results in (A.16)–(A.19) leads to

$$\begin{aligned} P(\max_{1\le k\le p}|\widehat{T}^*_{k3}-T_{k3}|\ge \widetilde{C}n^{-\kappa_2}/16) &\le 2p\exp\{-\widetilde{C}_5 n^{1-2\kappa_2-2\xi_3-2\xi_4}\} \\ &\quad + p\widetilde{C}_7 n^{\kappa_2}\exp(-8^{-1}c_0n^{\xi_3}) + \widetilde{C}_6 n^{\kappa_2+\xi_3}\exp\{-2^{-3/2}c_0n^{\xi_4/2}\}. \end{aligned}$$

Let $\xi_3 = (1-2\kappa_2)/3 - 2\eta$ and $\xi_4 = 3\eta$ with some $0 < \eta < (1-2\kappa_2)/6$. Then we have

$$P(\max_{1\le k\le p} |\widehat{T}^*_{k3} - T_{k3}| \ge \widetilde{C} n^{-\kappa_2}/16) \le p\widetilde{C}_8 \exp\{-\widetilde{C}_9 n^{(1-2\kappa_2)/3-2\eta}\} + \widetilde{C}_{10}\exp\{-\widetilde{C}_{11} n^{3\eta/2}\},$$

where $\widetilde{C}_8, \cdots, \widetilde{C}_{11}$ are some positive constants. This inequality together with (A.14)–(A.15) entails

$$P(\max_{1\le k\le p} |\widehat{T}_{k3} - T_{k3}| \ge \widetilde{C} n^{-\kappa_2}/4) \le p\widetilde{C}_1 \exp\{-\widetilde{C}_2 n^{(1-2\kappa_2)/3-2\eta}\} + \widetilde{C}_3\exp\{-\widetilde{C}_4 n^{3\eta/2}\} \tag{A.20}$$

for some positive constants $\widetilde{C}_1, \cdots, \widetilde{C}_4$.

D.4. Proof of Theorem 3. For simplicity, we provide here only the proof for the case without variable screening, that is, $d_1 = d_2 = p$. The case with variable screening can be proved using similar arguments, in view of the sure screening property established in Theorem 1. By the definition of $\widehat{\mathbf{B}}$, we have

$$\frac{1}{2nq}\|\mathbf{Y} - \widetilde{\mathbf{X}}\widehat{\mathbf{B}}\|_F^2 + \lambda\|\widehat{\mathbf{B}}\|_{2,1} \le \frac{1}{2nq}\|\mathbf{Y} - \widetilde{\mathbf{X}}\mathbf{B}^*\|_F^2 + \lambda\|\mathbf{B}^*\|_{2,1}.$$

Substituting $Y = \widetilde{\mathbf{X}}\mathbf{B}^* + \mathbf{W}$ and rearranging terms yield

$$\frac{1}{2nq}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F^2 \le \frac{1}{nq}\mathrm{tr}(\mathbf{W}^T\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}) + \lambda(\|\mathbf{B}^*\|_{2,1} - \|\widehat{\mathbf{B}}\|_{2,1}), \tag{A.21}$$

where $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{B}} - \mathbf{B}^*$. An application of the Cauchy-Schwarz inequality gives

$$\begin{aligned}\mathrm{tr}(\mathbf{W}^T\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}) = \mathrm{tr}(\widehat{\mathbf{\Delta}}\mathbf{W}^T\widetilde{\mathbf{X}}) &= \sum_k \widehat{\mathbf{\Delta}}_k \left[(\widetilde{\mathbf{X}}^T\mathbf{W})_k\right]^T \\ &\le \sum_k \|\widehat{\mathbf{\Delta}}_k\|_2 \cdot \|(\widetilde{\mathbf{X}}^T\mathbf{W})_k\|_2 \le \|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty}\|\widehat{\mathbf{\Delta}}\|_{2,1},\end{aligned} \tag{A.22}$$

where $\widehat{\mathbf{\Delta}}_k$ and $(\widetilde{\mathbf{X}}^T\mathbf{W})_k$ are the kth rows of $\widehat{\mathbf{\Delta}}$ and $\widetilde{\mathbf{X}}^T\mathbf{W}$, respectively.

Note that $q \le p$ and $\log p = o(n^\eta)$ with $\eta = \min\{\eta_0, 1/2 - 2\xi\}$. By Lemmas 7–8, with probability at least $1 - O\{\exp(-\widetilde{C}_1 n^{1/2-2\xi})\} - O(p^{-c}) = 1 - O(p^{-c_4})$ for some constants $\widetilde{C}_1, c, c_4 > 0$ it holds that

$$\frac{1}{nq}\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \le \frac{\lambda}{2} \tag{A.23}$$

and

$$\min_{|J|\le s,\, \mathbf{\Delta}\in\mathbb{R}^{\tilde{p}\times q}\setminus\{\mathbf{0}\},\, \|\mathbf{\Delta}_{J^c}\|_{2,1}\le 3\|\mathbf{\Delta}_J\|_{2,1}} \frac{\|\widetilde{\mathbf{X}}\mathbf{\Delta}\|_F}{\sqrt{n}\|\mathbf{\Delta}_J\|_F} \ge \frac{\kappa}{2}. \tag{A.24}$$

From now on, we condition on the event that these two inequalities hold. In view of (A.21) and (A.22), we have the following basic inequality

$$\begin{aligned}\frac{1}{2nq}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F^2 + \frac{\lambda}{2}\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} &\le \lambda(\|\mathbf{B}^*\|_{2,1} - \|\widehat{\mathbf{B}}\|_{2,1} \\ &\quad + \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1}) \le 2\lambda\|(\widehat{\mathbf{B}} - \mathbf{B}^*)_S\|_{2,1},\end{aligned} \tag{A.25}$$

where we have used the fact that $\|(\mathbf{B}^*)_{S^c}\|_{2,1} - \|(\widehat{\mathbf{B}})_{S^c}\|_{2,1} + \|(\widehat{\mathbf{B}} - \mathbf{B}^*)_{S^c}\|_{2,1} = 0$ and the triangle inequality.

The basic inequality (A.25) implies

$$\frac{1}{2nq}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F^2 \le 2\lambda\|\widehat{\mathbf{\Delta}}_S\|_{2,1} \le 2\lambda\sqrt{s}\|\widehat{\mathbf{\Delta}}_S\|_F, \tag{A.26}$$

where the last inequality holds since

$$\|\widehat{\mathbf{\Delta}}_S\|_{2,1} = \sum_{k\in S}\|\widehat{\mathbf{\Delta}}_k\|_2 \le \sqrt{s\sum_{k\in S}\|\widehat{\mathbf{\Delta}}_k\|_2^2} = \sqrt{s}\|\widehat{\mathbf{\Delta}}_S\|_F. \tag{A.27}$$

Moreover, it follows from (A.25) that $\|\widehat{\mathbf{\Delta}}_{S^c}\|_{2,1} \le 3\|\widehat{\mathbf{\Delta}}_S\|_{2,1}$ and thus by (A.24), we have $\|\widehat{\mathbf{\Delta}}_S\|_F \le 2\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F/(\kappa\sqrt{n})$. This inequality along with (A.25)–(A.27) yields

$$\frac{1}{2nq}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F^2 + \frac{\lambda}{2}\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \le \frac{4\lambda\sqrt{s}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F}{\kappa\sqrt{n}},$$

which gives $n^{-1/2}\|\widetilde{\mathbf{X}}\widehat{\mathbf{\Delta}}\|_F \le 8q\lambda\sqrt{s}/\kappa$. Therefore, we obtain

$$\frac{1}{2nq}\|\widetilde{\mathbf{X}}(\widehat{\mathbf{B}} - \mathbf{B}^*)\|_F^2 + \frac{\lambda}{2}\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{2,1} \le \frac{32qs\lambda^2}{\kappa^2},$$

which completes the proof for the first part of Theorem 3.

We next proceed to prove the second part of Theorem 3. We condition on the event that (A.4) holds. Denote by $S(\mathbf{B})$ the row support of any matrix $\mathbf{B}$. We need to show that with the same probability $S(\widetilde{\mathbf{B}}) = S(\mathbf{B}^*)$ holds. To this end, we first prove $S(\mathbf{B}^*) \subset S(\widetilde{\mathbf{B}})$. For any $j_0 \in S(\mathbf{B}^*)$, if $j_0 \notin S(\widetilde{\mathbf{B}})$ then the j_0th row of $\widetilde{\mathbf{B}}$ is zero, which means $\|\widehat{\mathbf{B}}_{j_0}\| \le 64c_3\kappa^{-2}s\sqrt{q(\log p)/n}$.

It then follows from the condition of $\min_{j\in S}\|\mathbf{B}_j^*\| > 128c_3\kappa^{-2}s\sqrt{q(\log p)/n}$ that

$$\begin{aligned}\frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}}-\mathbf{B}^*\|_{2,1} \geq & \frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}}_{j_0}-\mathbf{B}^*_{j_0}\| \geq \frac{1}{\sqrt{q}}(\|\mathbf{B}^*_{j_0}\|-\|\widehat{\mathbf{B}}_{j_0}\|)\\ >&64c_3\kappa^{-2}s\sqrt{(\log p)/n},\end{aligned}$$

which leads to a contradiction to the estimation bound (A.4). Thus it holds that $S(\mathbf{B}^*) \subset S(\widetilde{\mathbf{B}})$. We can also show that $S(\widetilde{\mathbf{B}}) \subset S(\mathbf{B}^*)$. In fact, if there exists some j_0 such that $j_0 \in S(\widetilde{\mathbf{B}})$ and $j_0 \notin S(\mathbf{B}^*)$, then we have $\|\widehat{\mathbf{B}}_{j_0}\| > 64c_3\kappa^{-2}s\sqrt{q(\log p)/n}$ and $\mathbf{B}^*_{j_0} = \mathbf{0}$, and thus

$$\frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}}-\mathbf{B}^*\|_{2,1} \geq \frac{1}{\sqrt{q}}\|\widehat{\mathbf{B}}_{j_0}-\mathbf{B}^*_{j_0}\| > 64c_3\kappa^{-2}s\sqrt{(\log p)/n},$$

which contradicts again the bound (A.4). Combining these results yields that with the same probability, the row support of $\widetilde{\mathbf{B}}$ is identical to the true row support S.

We finally prove (A.5). By assumption, the RE($2s$) condition holds. Using similar arguments as for proving (A.23)–(A.24), we can show that with probability at least $1 - O\{\exp(-\widetilde{C}_1 \cdot n^{1/2-2\xi})\} - O(p^{-c}) = 1 - O(p^{-c_4})$ for some constants $\widetilde{C}_1, c, c_4 > 0$, it holds that

$$\frac{1}{nq}\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \leq \frac{\lambda}{2} \tag{A.28}$$

and

$$\min_{|J|\leq 2s,\, \boldsymbol{\Delta}\in\mathbb{R}^{\tilde{p}\times q}\setminus\{\mathbf{0}\},\, \|\boldsymbol{\Delta}_{J^c}\|_{2,1}\leq 3\|\boldsymbol{\Delta}_J\|_{2,1}} \frac{\|\widetilde{\mathbf{X}}\boldsymbol{\Delta}\|_F}{\sqrt{n}\|\boldsymbol{\Delta}_J\|_F} \geq \frac{\kappa(2s)}{2}. \tag{A.29}$$

Recall that $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{B}} - \mathbf{B}^*$. Let S' be a subset of S^c corresponding to the s largest values of $\|\widehat{\boldsymbol{\Delta}}_k\|$. Then we have $|S \cup S'| = 2s$. From now on, we condition on the event that inequalities (A.28) and (A.29) hold. Conditional on such an event, the basic inequality (A.25) still holds. Thus we have $\|\widehat{\boldsymbol{\Delta}}_{S^c}\|_{2,1} \leq 3\|\widehat{\boldsymbol{\Delta}}_S\|_{2,1}$, which entails

$$\|\widehat{\boldsymbol{\Delta}}_{(S\cup S')^c}\|_{2,1} \leq \|\widehat{\boldsymbol{\Delta}}_{S^c}\|_{2,1} \leq 3\|\widehat{\boldsymbol{\Delta}}_S\|_{2,1} \leq 3\|\widehat{\boldsymbol{\Delta}}_{S\cup S'}\|_{2,1}.$$

This together with (A.29) yields $\|\widehat{\boldsymbol{\Delta}}_{S\cup S'}\|_F \leq 2\|\widetilde{\mathbf{X}}\widehat{\boldsymbol{\Delta}}\|_F/(\kappa(2s)\sqrt{n})$. From (A.26), we have

$$\frac{1}{2nq}\|\widetilde{\mathbf{X}}\widehat{\boldsymbol{\Delta}}\|_F^2 \leq 2\lambda\sqrt{s}\|\widehat{\boldsymbol{\Delta}}_S\|_F \leq 2\lambda\sqrt{s}\|\widehat{\boldsymbol{\Delta}}_{S\cup S'}\|_F.$$

Combining these two results gives

$$\|\widehat{\mathbf{\Delta}}_{S\cup S'}\|_F \le 16q\lambda\sqrt{s}/\kappa^2(2s). \tag{A.30}$$

Since the jth largest norm in the set $\{\|\widehat{\mathbf{\Delta}}_k\| : k \in S^c\}$ is bounded from above by $\|\widehat{\mathbf{\Delta}}_{S^c}\|_{2,1}/j$, it holds that

$$\begin{aligned}\sum_{k\in(S\cup S')^c}\|\widehat{\mathbf{\Delta}}_k\|^2 &\le \sum_{k=s+1}^{\widetilde{p}-s}\frac{\|\widehat{\mathbf{\Delta}}_{S^c}\|_{2,1}^2}{k^2} \le \frac{\|\widehat{\mathbf{\Delta}}_{S^c}\|_{2,1}^2}{s} \le \frac{9\|\widehat{\mathbf{\Delta}}_S\|_{2,1}^2}{s}\\ &\le 9\sum_{k\in S}\|\widehat{\mathbf{\Delta}}_k\|^2 \le 9\sum_{k\in S\cup S'}\|\widehat{\mathbf{\Delta}}_k\|^2,\end{aligned}$$

which results in $\|\widehat{\mathbf{\Delta}}\|_F^2 \le 10\|\widehat{\mathbf{\Delta}}_{S\cup S'}\|_F^2$. This inequality along with (A.30) yields

$$\frac{1}{\sqrt{q}}\|\widehat{\mathbf{\Delta}}\|_F \le \frac{16\sqrt{10}c_3}{\kappa^2(2s)}\sqrt{s(\log p)/n},$$

which concludes the proof for the third part of Theorem 3.

APPENDIX E: ADDITIONAL TECHNICAL DETAILS AND LEMMAS

E.1. Terms $E(Y|X_j)$ and $E(Y^2|X_j)$ under model (2). Since the covariates $X_1,\cdots,X_p$ are all independent with mean zero and the random error W is of mean zero and independent of all X_j's, it is immediate that $E(Y|X_j) = \alpha + \beta_j X_j$. We now calculate $E(Y^2|X_j)$. Define

$$\begin{aligned}&J_1 = \sum_{j=1}^{p}\beta_j X_j, \quad J_2 = \sum_{k=1}^{p-1}\sum_{\ell=k+1}^{p}\gamma_{k\ell}X_kX_\ell, \quad J_3 = \sum_{k\neq j}\beta_k X_k,\\ &J_4 = \sum_{k=1}^{j-1}\gamma_{kj}X_k + \sum_{\ell=j+1}^{p}\gamma_{j\ell}X_\ell, \quad J_5 = \sum_{k=1,k\neq j}^{p-1}\sum_{\ell=k+1,\ell\neq j}^{p}\gamma_{k\ell}X_kX_\ell.\end{aligned}$$

Then we have $Y = \alpha + J_1 + J_2 + W$ with $J_1 = \beta_j X_j + J_3$ and $J_2 = J_4X_j + J_5$, and J_3, J_4, and J_5 are independent of X_j. Applying the properties of conditional expectation yields

$$\begin{aligned}E[(\alpha + J_1 + J_2)W|X_j] &= E\{E[(\alpha + J_1 + J_2)W|X_1,\cdots,X_p]|X_j\}\\ &= E[(\alpha + J_1 + J_2)E(W|X_1,\cdots,X_p)|X_j] = 0\end{aligned}$$

and

$$
\begin{aligned}
&E[(\alpha+J_1+J_2)^2|X_j] = E\{[(\beta_j+J_4)X_j+(\alpha+J_3+J_5)]^2|X_j\}\\
&=X_j^2E[(\beta_j+J_4)^2]+2X_jE[(\beta_j+J_4)(\alpha+J_3+J_5)]+E[(\alpha+J_3+J_5)^2]\\
&=\left[\beta_j^2+\sum_{k=1}^{j-1}\gamma_{kj}^2E(X_k^2)+\sum_{\ell=j+1}^{p}\gamma_{j\ell}^2E(X_\ell^2)\right]X_j^2\\
&\quad+2\left[\beta_j\alpha+\sum_{k=1}^{j-1}\beta_k\gamma_{kj}E(X_k^2)+\sum_{\ell=j+1}^{p}\beta_\ell\gamma_{j\ell}E(X_\ell^2)\right]X_j\\
&\quad+\alpha^2+\sum_{k\neq j}\beta_k^2E(X_k^2)+\sum_{k=1,k\neq j}^{p-1}\sum_{\ell=k+1,\ell\neq j}^{p}\gamma_{k\ell}^2E(X_k^2)E(X_\ell^2).
\end{aligned}
$$

Therefore, it holds that

$$
\begin{aligned}
E(Y^2|X_j) &= E[(\alpha+J_1+J_2)^2|X_j]+2E[(\alpha+J_1+J_2)W|X_j]+E(W^2|X_j)\\
&=\left[\beta_j^2+\sum_{k=1}^{j-1}\gamma_{kj}^2E(X_k^2)+\sum_{\ell=j+1}^{p}\gamma_{j\ell}^2E(X_\ell^2)\right]X_j^2\\
&\quad+2\left[\beta_j\alpha+\sum_{k=1}^{j-1}\beta_k\gamma_{kj}E(X_k^2)+\sum_{\ell=j+1}^{p}\beta_\ell\gamma_{j\ell}E(X_\ell^2)\right]X_j+C_j,
\end{aligned}
$$

where $C_j=\alpha^2+\sum_{k\neq j}\beta_k^2E(X_k^2)+\sum_{k=1,k\neq j}^{p-1}\sum_{\ell=k+1,\ell\neq j}^{p}\gamma_{k\ell}^2E(X_k^2)E(X_\ell^2)+\sigma^2$ is a constant that is free of X_j, and σ^2 is the variance of W.

E.2. Lemma 1 and its proof.

LEMMA 1. *Let $\widehat{A}$ and $\widehat{B}$ be estimates of A and B, respectively, based on a sample of size n. Assume that both A and B are bounded and for any constant $\widetilde{C}>0$, there exist positive constants $\widetilde{C}_1,\cdots,\widetilde{C}_4$ such that*

$$
\begin{aligned}
P\left(|\widehat{A}-A|\geq\widetilde{C}n^{-\kappa}\right)&\leq\widetilde{C}_1\exp\left\{-\widetilde{C}_2n^{f(\kappa)}\right\}\\
P\left(|\widehat{B}-B|\geq\widetilde{C}n^{-\kappa}\right)&\leq\widetilde{C}_3\exp\left\{-\widetilde{C}_4n^{f(\kappa)}\right\}
\end{aligned}
$$

with $f(\kappa)$ some function of κ. Then for any constant $\widetilde{C}>0$, there exist positive constants $\widetilde{C}_5$ and $\widetilde{C}_6$ such that

$$
P(|\widehat{A}\widehat{B}-AB|\geq\widetilde{C}n^{-\kappa})\leq\widetilde{C}_5\exp\left\{-\widetilde{C}_6n^{f(\kappa)}\right\}.
$$

Proof. Note that $|\widehat{A}\widehat{B} - AB| \le |\widehat{A}(\widehat{B} - B)| + |(\widehat{A} - A)B|$. Thus for any positive constant $\widetilde{C}$, we have

$$P(|\widehat{A}\widehat{B} - AB| \ge \widetilde{C}n^{-\kappa}) \le P(|\widehat{A}(\widehat{B} - B)| \ge \widetilde{C}n^{-\kappa}/2) + P(|(\widehat{A} - A)B| \ge \widetilde{C}n^{-\kappa}/2). \tag{A.31}$$

We first deal with the second term on the right hand side of (A.31). Since both A and B are bounded, there exists some positive constant L such that $|A| \le L$ and $|B| \le L$. It follows that

$$\begin{aligned} P(|(\widehat{A} - A)B| \ge \widetilde{C}n^{-\kappa}/2) &\le P(|\widehat{A} - A|L \ge \widetilde{C}n^{-\kappa}/2) \\ &= P\{|\widehat{A} - A| \ge (2L)^{-1}\widetilde{C}n^{-\kappa_1}\} \le \widetilde{C}_1 \exp\left\{-\widetilde{C}_2 n^{f(\kappa_1)}\right\}, \end{aligned} \tag{A.32}$$

where $\widetilde{C}_1$ and $\widetilde{C}_2$ are some positive constants.

We next consider the first term on the right hand side of (A.31). Note that

$$\begin{aligned} P(|\widehat{A}(\widehat{B} - B)| \ge \widetilde{C}n^{-\kappa}/2) &\le P\Big\{|\widehat{A}(\widehat{B} - B)| \ge \widetilde{C}n^{-\kappa}/2, \\ &\quad |\widehat{A}| \ge L + \frac{\widetilde{C}}{2}n^{-\kappa}\Big\} + P\Big(|\widehat{A}(\widehat{B} - B)| \ge \frac{\widetilde{C}}{2}n^{-\kappa}, |\widehat{A}| < L + \frac{\widetilde{C}}{2}n^{-\kappa}\Big) \\ &\le P(|\widehat{A}| \ge L + \frac{\widetilde{C}}{2}n^{-\kappa}) + P(|\widehat{A}(\widehat{B} - B)| \ge \frac{\widetilde{C}}{2}n^{-\kappa}, |\widehat{A}| < L + \widetilde{C}) \\ &\le P(|\widehat{A}| \ge L + \widetilde{C}n^{-\kappa}/2) + P\{(L + \widetilde{C})|\widehat{B} - B| \ge \widetilde{C}n^{-\kappa}/2\}. \end{aligned} \tag{A.33}$$

We will bound the two terms on the right hand side of (A.33) separately. It follows from $|A| \le L$ that

$$\begin{aligned} P(|\widehat{A}| \ge L + \widetilde{C}n^{-\kappa}/2) &\le P(|\widehat{A} - A| + |A| \ge L + \widetilde{C}n^{-\kappa}/2) \\ &\le P\{|\widehat{A} - A| \ge 2^{-1}\widetilde{C}n^{-\kappa}\} \le \widetilde{C}_3 \exp\left\{-\widetilde{C}_4 n^{f(\kappa_1)}\right\}, \end{aligned} \tag{A.34}$$

where $\widetilde{C}_3$ and $\widetilde{C}_4$ are some positive constants. It also holds that

$$\begin{aligned} P((L + \widetilde{C})|\widehat{B} - B| \ge \widetilde{C}n^{-\kappa}/2) &= P\{|\widehat{B} - B| \ge (2L + 2\widetilde{C})^{-1}\widetilde{C}n^{-\kappa}\} \\ &\le \widetilde{C}_7 \exp\left\{-\widetilde{C}_8 n^{f(\kappa)}\right\}, \end{aligned}$$

where $\widetilde{C}_7$ and $\widetilde{C}_8$ are some positive constants. This inequality together with (A.31)–(A.34) entails

$$\begin{aligned} P(|\widehat{A}\widehat{B} - AB| \ge \widetilde{C}n^{-\kappa}) &\le \widetilde{C}_1 \exp\left\{-\widetilde{C}_2 n^{f(\kappa)}\right\} + \widetilde{C}_3 \exp\left\{-\widetilde{C}_4 n^{f(\kappa)}\right\} \\ &\quad + \widetilde{C}_7 \exp\left\{-\widetilde{C}_8 n^{f(\kappa)}\right\} \le \widetilde{C}_5 \exp\left\{-\widetilde{C}_6 n^{f(\kappa)}\right\}, \end{aligned}$$

where $\widetilde{C}_5 = \widetilde{C}_1 + \widetilde{C}_3 + \widetilde{C}_7$ and $\widetilde{C}_6 = \min\{\widetilde{C}_2, \widetilde{C}_4, \widetilde{C}_8\}$.

E.3. Lemma 2 and its proof. For any set $\mathcal{D}$, we denote by $|\mathcal{D}|$ its cardinality throughout the paper.

LEMMA 2. *Let $\widehat{B}_j \geq 0$ be an estimate of B_j based on a sample of size n for each $j \in \mathcal{D} \subset \{1, \cdots, p\}$. Assume that $\min_{j\in\mathcal{D}} B_j \geq L$ for some positive constant L, and for any constant $\widetilde{C} > 0$, there exist positive constants $\widetilde{C}_1$ and $\widetilde{C}_2$ such that*

$$P\left(\max_{j\in\mathcal{D}} |\widehat{B}_j - B_j| \geq \widetilde{C} n^{-\kappa}\right) \leq |\mathcal{D}|\widetilde{C}_1 \exp\left\{-\widetilde{C}_2 n^{f(\kappa)}\right\}$$

with $f(\kappa)$ some function of κ. Then for any constant $\widetilde{C} > 0$, there exist positve constants $\widetilde{C}_3$ and $\widetilde{C}_4$ such that

$$P\left(\max_{j\in\mathcal{D}} |\sqrt{\widehat{B}_j} - \sqrt{B_j}| \geq \widetilde{C} n^{-\kappa}\right) \leq |\mathcal{D}|\widetilde{C}_3 \exp\left\{-\widetilde{C}_4 n^{f(\kappa)}\right\}.$$

Proof. Since $\min_{j\in\mathcal{D}} B_j \geq L$ for some positive constant L, there exists a constant L_0 such that $0 < L_0 < L$. Note that for any positive constant $\widetilde{C}$,

$$\begin{aligned}
P(\max_{j\in\mathcal{D}} |\sqrt{\widehat{B}_j} - \sqrt{B_j}| \geq \widetilde{C} n^{-\kappa}) &\leq P\Big\{ \max_{j\in\mathcal{D}} |\sqrt{\widehat{B}_j} - \sqrt{B_j}| \geq \widetilde{C} n^{-\kappa}, \\
&\min_{j\in\mathcal{D}} |\widehat{B}_j| \leq L - L_0 n^{-\kappa}\Big\} + P\Big\{ \max_{j\in\mathcal{D}} |\sqrt{\widehat{B}_j} - \sqrt{B_j}| \geq \widetilde{C} n^{-\kappa}, \\
&\min_{j\in\mathcal{D}} |\widehat{B}_j| > L - L_0 n^{-\kappa}\Big\} \leq P(\min_{j\in\mathcal{D}} |\widehat{B}_j| \leq L - L_0 n^{-\kappa}) \\
&+ P(\max_{j\in\mathcal{D}} \frac{|\widehat{B}_j - B_j|}{|\sqrt{\widehat{B}_j} + \sqrt{B_j}|} \geq \widetilde{C} n^{-\kappa}, \min_{j\in\mathcal{D}} |\widehat{B}_j| > L - L_0).
\end{aligned} \tag{A.35}$$

We first consider the first term on the right hand side of (A.35). It follows from $\min_{j\in\mathcal{D}} B_j \geq L$ that

$$\begin{aligned}
P(\min_{j\in\mathcal{D}} |\widehat{B}_j| \leq L - L_0 n^{-\kappa}) &\leq P\Big\{ \min_{j\in\mathcal{D}} |B_j| - \max_{j\in\mathcal{D}} |\widehat{B}_j - B_j| \\
&\leq L - L_0 n^{-\kappa}\Big\} \leq P(\max_{j\in\mathcal{D}} |\widehat{B}_j - B_j| \geq L_0 n^{-\kappa}) \\
&\leq |\mathcal{D}|\widetilde{C}_1 \exp\left\{-\widetilde{C}_2 n^{f(\kappa)}\right\},
\end{aligned} \tag{A.36}$$

where $\widetilde{C}_1$ and $\widetilde{C}_2$ are some positive constants.

Next we consider the second term on the right hand side of (A.35). For any positive constant $\widetilde{C}$, we have

$$
\begin{aligned}
(\text{A.37}) \qquad & P(\max_{j\in\mathcal{D}} \frac{|\widehat{B}_j - B_j|}{|\sqrt{\widehat{B}_j} + \sqrt{B_j}|} \geq \widetilde{C} n^{-\kappa}, \min_{j\in\mathcal{D}} |\widehat{B}_j| > L - L_0) \\
& \leq P\{\max_{j\in\mathcal{D}} |\widehat{B}_j - B_j| \geq \widetilde{C}(\sqrt{L - L_0} + \sqrt{L}) n^{-\kappa}\} \\
& \leq |\mathcal{D}| \widetilde{C}_5 \exp\left\{-\widetilde{C}_6 n^{f(\kappa)}\right\},
\end{aligned}
$$

where $\widetilde{C}_5$ and $\widetilde{C}_6$ are some positive constants. Combining (A.35)–(A.37) yields

$$
(\text{A.38}) \qquad P(\max_{j\in\mathcal{D}} |\sqrt{\widehat{B}_j} - \sqrt{B_j}| \geq \widetilde{C} n^{-\kappa}) \leq |\mathcal{D}| \widetilde{C}_3 \exp\left\{-\widetilde{C}_4 n^{f(\kappa)}\right\},
$$

where $\widetilde{C}_3 = \widetilde{C}_1 + \widetilde{C}_5$ and $\widetilde{C}_4 = \min\{\widetilde{C}_2, \widetilde{C}_6\}$.

E.4. Lemma 3 and its proof.

LEMMA 3. *Let $\widehat{A}_j$ and $\widehat{B}_j$ be estimates of A_j and B_j, respectively, based on a sample of size n for each $j \in \mathcal{D} \subset \{1, \cdots, p\}$. Assume that A_j and B_j satisfy $\max_{j\in\mathcal{D}} |A_j| \leq L_1$ and $\min_{j\in\mathcal{D}} |B_j| \geq L_2$ for some constants $L_1, L_2 > 0$, and for any constant $\widetilde{C} > 0$, there exist constants $\widetilde{C}_1, \cdots, \widetilde{C}_6 > 0$ such that*

$$
\begin{aligned}
& P\big(\max_{j\in\mathcal{D}} |\widehat{A}_j - A_j| \geq \widetilde{C} n^{-\kappa}\big) \leq |\mathcal{D}| \widetilde{C}_1 \exp\big\{-\widetilde{C}_2 n^{f(\kappa)}\big\} + \widetilde{C}_3 \exp\big\{-\widetilde{C}_4 n^{g(\kappa)}\big\}, \\
& P\big(\max_{j\in\mathcal{D}} |\widehat{B}_j - B_j| \geq \widetilde{C} n^{-\kappa}\big) \leq |\mathcal{D}| \widetilde{C}_5 \exp\big\{-\widetilde{C}_6 n^{f(\kappa)}\big\}
\end{aligned}
$$

with $f(\kappa)$ and $g(\kappa)$ some functions of κ. Then for any constant $\widetilde{C} > 0$, there exist positive constants $\widetilde{C}_7, \cdots, \widetilde{C}_{10}$ such that

$$
\begin{aligned}
P\left(\max_{j\in\mathcal{D}} \left|\frac{\widehat{A}_j}{\widehat{B}_j} - \frac{A_j}{B_j}\right| \geq \widetilde{C} n^{-\kappa}\right) & \leq |\mathcal{D}| \widetilde{C}_7 \exp\left\{-\widetilde{C}_8 n^{f(\kappa)}\right\} \\
& \quad + \widetilde{C}_9 \exp\left\{-\widetilde{C}_{10} n^{g(\kappa)}\right\}.
\end{aligned}
$$

Proof. Since $\min_{j\in\mathcal{D}} |B_j| \geq L_2 > 0$, there exists some constant L_0 such

that $0 < L_0 < L_2$. Note that for any positive constant $\widetilde{C}$, we have

$$
\begin{aligned}
(\text{A.39}) \qquad P(\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa}) &\leq P\Big\{\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa},\\
&\min_{j\in\mathcal{D}}|\widehat{B}_j|\leq L_2-L_0n^{-\kappa}\Big\}+P\Big\{\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa},\\
&\min_{j\in\mathcal{D}}|\widehat{B}_j|> L_2-L_0n^{-\kappa}\Big\}\leq P(\min_{j\in\mathcal{D}}|\widehat{B}_j|\leq L_2-L_0n^{-\kappa})\\
&+P(\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa},\min_{j\in\mathcal{D}}|\widehat{B}_j|> L_2-L_0).
\end{aligned}
$$

We start with the first term on the right hand side of (A.39). In light of $\min_{j\in\mathcal{D}}|B_j|\geq L_2$, we deduce

$$
\begin{aligned}
(\text{A.40}) \qquad P(\min_{j\in\mathcal{D}}|\widehat{B}_j|\leq L_2-L_0n^{-k}) &\leq P\Big\{\min_{j\in\mathcal{D}}|B_j|-\max_{j\in\mathcal{D}}|\widehat{B}_j-B_j|\\
&\leq L_2-L_0n^{-\kappa}\Big\}\leq P(\max_{j\in\mathcal{D}}|\widehat{B}_j-B_j|\geq L_0n^{-\kappa})\\
&\leq |\mathcal{D}|\widetilde{C}_1\exp\left\{-\widetilde{C}_2n^{f(\kappa)}\right\},
\end{aligned}
$$

where $\widetilde{C}_1$ and $\widetilde{C}_2$ are some positive constants.

The second term on the right hand side of (A.39) can be bounded as

$$
\begin{aligned}
(\text{A.41}) \qquad & P(\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa},\min_{j\in\mathcal{D}}|\widehat{B}_j|> L_2-L_0)\\
&\leq P(\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{\widehat{B}_j}|\geq \widetilde{C}n^{-\kappa}/2,\min_{j\in\mathcal{D}}|\widehat{B}_j|> L_2-L_0)\\
&\quad+P(\max_{j\in\mathcal{D}}|\frac{A_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa}/2,\min_{j\in\mathcal{D}}|\widehat{B}_j|> L_2-L_0)\\
&\leq P\{\max_{j\in\mathcal{D}}|\widehat{A}_j-A_j|\geq 2^{-1}(L_2-L_0)\widetilde{C}n^{-\kappa}\}\\
&\quad+P\{\max_{j\in\mathcal{D}}|\widehat{B}_j-B_j|\geq (2L_1)^{-1}(L_2-L_0)L_2\widetilde{C}n^{-\kappa}\}\\
&\leq |\mathcal{D}|\widetilde{C}_3\exp\left\{-\widetilde{C}_4n^{f(\kappa)}\right\}+\widetilde{C}_9\exp\left\{-\widetilde{C}_{10}n^{g(\kappa)}\right\}\\
&\quad+|\mathcal{D}|\widetilde{C}_5\exp\left\{-\widetilde{C}_6n^{f(\kappa)}\right\},
\end{aligned}
$$

where $\widetilde{C}_3,\cdots,\widetilde{C}_6$, and $\widetilde{C}_9,\widetilde{C}_{10}$ are some positive constants. Combining (A.39)–

(A.41) results in

$$P(\max_{j\in\mathcal{D}}|\frac{\widehat{A}_j}{\widehat{B}_j}-\frac{A_j}{B_j}|\geq \widetilde{C}n^{-\kappa})\leq |\mathcal{D}|\widetilde{C}_7\exp\left\{-\widetilde{C}_8 n^{f(\kappa)}\right\} + \widetilde{C}_9\exp\left\{-\widetilde{C}_{10}n^{g(\kappa)}\right\},$$

where $\widetilde{C}_7=\widetilde{C}_1+\widetilde{C}_3+\widetilde{C}_5$ and $\widetilde{C}_8=\min\{\widetilde{C}_2,\widetilde{C}_4,\widetilde{C}_6\}$.

E.5. Lemma 4 and its proof.

LEMMA 4. *Let Z be a nonnegative random variable satisfying $P(Z>t)\leq \widetilde{C}_1\exp(-\widetilde{C}_2t^2)$ for all $t>0$ with $\widetilde{C}_1,\widetilde{C}_2>0$ some constants. Then*

$$E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z^2\right)\right]\leq 1+\widetilde{C}_1 \quad \textit{and} \quad E(Z^{2m})\leq (1+\widetilde{C}_1)(2\widetilde{C}_2^{-1})^m m!$$

for any nonnegative integer m.

Proof. Let $F(t)$ be the cumulative distribution function of Z. Then

$$1-F(t)=P(Z>t)\leq \widetilde{C}_1\exp(-\widetilde{C}_2t^2)$$

for all $t>0$. Using integration by parts, we have

$$\begin{aligned}E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z^2\right)\right] &= -\int_0^\infty \exp\left(\frac{\widetilde{C}_2}{2}t^2\right)\,d[1-F(t)]\\ &= 1+\int_0^\infty \widetilde{C}_2 t\exp\left(\frac{\widetilde{C}_2}{2}t^2\right)[1-F(t)]\,dt\\ &\leq 1+\widetilde{C}_1\int_0^\infty c_2 t\exp\left(-\frac{\widetilde{C}_2}{2}t^2\right)\,dt=1+\widetilde{C}_1.\end{aligned}$$

With the Taylor series of the exponential function, we obtain

$$E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z^2\right)\right]=\sum_{k=0}^\infty \frac{\widetilde{C}_2^k E(Z^{2k})}{2^k k!}\geq \frac{\widetilde{C}_2^m E(Z^{2m})}{2^m m!}$$

for any nonnegative integer m. Thus $E(Z^{2m})\leq (1+\widetilde{C}_1)(2\widetilde{C}_2^{-1})^m m!$.

E.6. Lemma 5 and its proof.

LEMMA 5. *Let Z be a nonnegative random variable satisfying $P(Z > t) \le \widetilde{C}_1 \exp(-\widetilde{C}_2 t)$ for all $t > 0$ with $\widetilde{C}_1, \widetilde{C}_2 > 0$ some constants. Then*

$$E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z\right)\right] \le 1 + \widetilde{C}_1 \quad \textit{and} \quad E(Z^m) \le (1+\widetilde{C}_1)(2\widetilde{C}_2^{-1})^m m!$$

for any nonnegative integer m.

Proof. Let $F(t)$ be the cumulative distribution function of Z. Then

$$1 - F(t) = P(Z > t) \le \widetilde{C}_1 \exp(-\widetilde{C}_2 t)$$

for all $t > 0$. It follows from integration by parts that

$$\begin{aligned} E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z\right)\right] &= -\int_0^\infty \exp\left(\frac{\widetilde{C}_2}{2}t\right) d[1-F(t)] \\ &= 1 + \int_0^\infty \exp\left(\frac{\widetilde{C}_2}{2}t\right)[1-F(t)]\,dt \\ &\le 1 + \widetilde{C}_1 \int_0^\infty \frac{\widetilde{C}_2}{2}\exp\left(-\frac{\widetilde{C}_2}{2}t\right) dt = 1 + \widetilde{C}_1. \end{aligned}$$

Applying the Taylor series of the exponential function leads to

$$E\left[\exp\left(\frac{\widetilde{C}_2}{2}Z\right)\right] = \sum_{k=0}^\infty \frac{\widetilde{C}_2^k E(Z^k)}{2^k k!} \ge \frac{\widetilde{C}_2^m E(Z^m)}{2^m m!}$$

for any nonnegative integer m. Thus $E(Z^m) \le (1+\widetilde{C}_1)(2\widetilde{C}_2^{-1})^m m!$.

E.7. Lemma 6 and its proof.

LEMMA 6. *Under Condition 2, both $dcov^2(X_k, \mathbf{y})$ and $dcov^2(X_k^*, \mathbf{y}^*)$ are uniformly bounded in k.*

Proof. We will show that $\mathrm{dcov}^2(X_k^*, \mathbf{y}^*)$ are uniformly bounded in $1 \le k \le p$. Similar arguments apply to prove that $\mathrm{dcov}^2(X_k, \mathbf{y})$ are also uniformly bounded in k. Recall that $\mathrm{dcov}^2(X_k^*, \mathbf{y}^*) = T_{k1} + T_{k2} - 2T_{k3}$ where $T_{k1} = E\left[\phi(X_{1k}^*, X_{2k}^*)\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)\right]$, $T_{k2} = E\left[\phi(X_{1k}^*, X_{2k}^*)\right] E\left[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)\right]$, and $T_{k3} =$

$E[\phi(X_{1k}^*, X_{2k}^*)\psi(\mathbf{y}_1^*, \mathbf{y}_3^*)]$. Here $\phi(X_{1k}^*, X_{2k}^*) = |X_{1k}^* - X_{2k}^*|$ and $\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) = \|\mathbf{y}_1^* - \mathbf{y}_2^*\|$. Thus an application of the triangle inequality gives

$$0 \le \mathrm{dcov}^2(X_k^*, \mathbf{y}^*) \le |T_{k1}| + |T_{k2}| + 2|T_{k3}|. \tag{A.42}$$

To prove that $\mathrm{dcov}^2(X_k^*, \mathbf{y}^*)$ are uniformly bounded in k, it suffices to show that each term on the right hand side above is uniformly bounded in view of (A.42). As shown in Step 1 of Part 1 in the proof of Theorem 1, under Condition 2 the first quantity T_{k1} is uniformly bounded in $1 \le k \le p$. Using similar arguments, we can show that T_{k2} and T_{k3} are also uniformly bounded in k, which completes the proof.

E.8. Lemma 7 and its proof.

LEMMA 7. *Assume that Conditions 4–5 hold and* $\log p = o(n^{1/2-2\xi})$. *Then with probability at least* $1 - O\{\exp(-\widetilde{C}_1 n^{1/2-2\xi})\}$ *for some constant* $\widetilde{C}_1 > 0$, *it holds that*

$$\min_{|J| \le s,\, \mathbf{\Delta} \in \mathbb{R}^{\widetilde{p} \times q} \setminus \{\mathbf{0}\},\, \|\mathbf{\Delta}_{J^c}\|_{2,1} \le 3\|\mathbf{\Delta}_J\|_{2,1}} \frac{\|\widetilde{\mathbf{X}}\mathbf{\Delta}\|_F}{\sqrt{n}\|\mathbf{\Delta}_J\|_F} \ge \frac{\kappa}{2}.$$

Proof. The main idea of the proof is to first introduce an event with a high probability and then derive the desired inequality conditional on that event. Define an event

$$\mathcal{E} = \{\|n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} - \mathbf{\Sigma}\|_\infty < \epsilon\},$$

where $\|\cdot\|_\infty$ denotes the entrywise matrix L_∞-norm and $0 < \epsilon < 1$ will be specified later. In view of the first part of Condition 4, it follows from Lemmas 4 and 10 that $P(\mathcal{E}) \ge 1 - \widetilde{C}_2\widetilde{p}^2\exp(-\widetilde{C}_3 n^{1/2}\epsilon^2)$ for some constants $\widetilde{C}_2, \widetilde{C}_3 > 0$.

From now on, we condition on the event $\mathcal{E}$. By the definition of the Frobenius norm, we have

$$n^{-1}\|\widetilde{\mathbf{X}}\mathbf{\Delta}\|_F^2 = \mathrm{tr}[\mathbf{\Delta}^T(n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} - \mathbf{\Sigma})\mathbf{\Delta}] + \mathrm{tr}(\mathbf{\Delta}^T\mathbf{\Sigma}\mathbf{\Delta}). \tag{A.43}$$

For any matrix $\mathbf{M}$, denote by $\mathbf{M}_{ij}$ the (i,j)-entry of $\mathbf{M}$. Then conditional on the event $\mathcal{E}$, the first term on the right hand of (A.43) can be bounded

as

$$\begin{aligned}(\text{A.44}) \quad \left|\mathrm{tr}[\boldsymbol{\Delta}^T(n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}-\boldsymbol{\Sigma})\boldsymbol{\Delta}]\right| &= \left|\mathrm{tr}[(n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}-\boldsymbol{\Sigma})\boldsymbol{\Delta}\boldsymbol{\Delta}^T]\right| \\ &= |\sum_{i=1}^{\widetilde{p}}\sum_{j=1}^{\widetilde{p}}(n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}-\boldsymbol{\Sigma})_{ij}(\boldsymbol{\Delta}\boldsymbol{\Delta}^T)_{ij}| \\ &\le \epsilon\sum_{i=1}^{\widetilde{p}}\sum_{j=1}^{\widetilde{p}}\sum_{k=1}^{q}|\boldsymbol{\Delta}_{ik}||\boldsymbol{\Delta}_{jk}| = \epsilon\sum_{k=1}^{q}\left(\sum_{j\in J}|\boldsymbol{\Delta}_{jk}|+\sum_{j\in J^c}|\boldsymbol{\Delta}_{jk}|\right)^2 \\ &\le 2\epsilon\sum_{k=1}^{q}\left(\sum_{j\in J}|\boldsymbol{\Delta}_{jk}|\right)^2 + 2\epsilon\sum_{k=1}^{q}\left(\sum_{j\in J^c}|\boldsymbol{\Delta}_{jk}|\right)^2,\end{aligned}$$

where we have used the fact that $(a+b)^2 \le 2(a^2+b^2)$ in the last inequality.

By the Cauchy-Schwarz inequality, for any set J satisfying $|J| \le s$ we have

$$(\text{A.45}) \quad \sum_{k=1}^{q}\left(\sum_{j\in J}|\boldsymbol{\Delta}_{jk}|\right)^2 \le \sum_{k=1}^{q}|J|\sum_{j\in J}\boldsymbol{\Delta}_{jk}^2 = |J|\cdot\|\boldsymbol{\Delta}_J\|_F^2 \le s\|\boldsymbol{\Delta}_J\|_F^2.$$

For any $\boldsymbol{\Delta} \in \mathbb{R}^{\widetilde{p}\times q}\backslash\{\mathbf{0}\}$ satisfying $\|\boldsymbol{\Delta}_{J^c}\|_{2,1} \le 3\|\boldsymbol{\Delta}_J\|_{2,1}$ with $|J| \le s$, similar arguments apply to show that

$$\begin{aligned}&\sum_{k=1}^{q}\left(\sum_{j\in J^c}|\boldsymbol{\Delta}_{jk}|\right)^2 = \sum_{j\in J^c}\sum_{j'\in J^c}\sum_{k=1}^{q}|\boldsymbol{\Delta}_{jk}||\boldsymbol{\Delta}_{j'k}| \le \sum_{j\in J^c}\sum_{j'\in J^c}\left(\sum_{k=1}^{q}\boldsymbol{\Delta}_{jk}^2\right)^{1/2} \\ &\quad\cdot\left(\sum_{k=1}^{q}\boldsymbol{\Delta}_{j'k}^2\right)^{1/2} = \sum_{j\in J^c}\sum_{j'\in J^c}\|\boldsymbol{\Delta}_j\|_2\cdot\|\boldsymbol{\Delta}_{j'}\|_2 = \left(\sum_{j\in J^c}\|\boldsymbol{\Delta}_j\|_2\right)^2 \\ &= \|\boldsymbol{\Delta}_{J^c}\|_{2,1}^2 \le 9\|\boldsymbol{\Delta}_J\|_{2,1}^2 = 9\left(\sum_{j\in J}\|\boldsymbol{\Delta}_j\|_2\right)^2 \le 9|J|\left(\sum_{j\in J}\|\boldsymbol{\Delta}_j\|_2^2\right) \\ &= 9|J|\|\boldsymbol{\Delta}_J\|_F^2 \le 9s\|\boldsymbol{\Delta}_J\|_F^2,\end{aligned}$$

which along with (A.44)–(A.45) entails

$$|\mathrm{tr}[\boldsymbol{\Delta}^T(n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}-\boldsymbol{\Sigma})\boldsymbol{\Delta}]| \le 20s\epsilon\|\boldsymbol{\Delta}_J\|_F^2.$$

Combining the above inequality with (A.43) and by Condition 5, we obtain

$$\min_{|J|\le s,\mathbf{\Delta}\in\mathbb{R}^{\widetilde{p}\times q}\setminus\{\mathbf{0}\},\|\mathbf{\Delta}_{J^c}\|_{2,1}\le 3\|\mathbf{\Delta}_J\|_{2,1}} \left(\frac{\|\widetilde{\mathbf{X}}\mathbf{\Delta}\|_F}{\sqrt{n}\|\mathbf{\Delta}_J\|_F}\right)^2 \ge \kappa^2 - 20s\epsilon.$$

It follows from the second part of Condition that $s \le \widetilde{C}_4 n^{\xi}$ for some positive constant $\widetilde{C}_4$. We choose $\epsilon = 3\kappa^2/(80\widetilde{C}_4 n^{\xi})$. Then we have $\kappa^2 - 20s\epsilon \ge \kappa^2/4$ and for sufficiently large n, $0 < \epsilon < 1$. Therefore, it holds with probability at least $1 - O\{\exp(-\widetilde{C}_1 n^{1/2-2\xi})\}$ for some constant $\widetilde{C}_1 > 0$ that

$$\min_{|J|\le s,\mathbf{\Delta}\in\mathbb{R}^{\widetilde{p}\times q}\setminus\{\mathbf{0}\},\|\mathbf{\Delta}_{J^c}\|_{2,1}\le 3\|\mathbf{\Delta}_J\|_{2,1}} \frac{\|\widetilde{\mathbf{X}}\mathbf{\Delta}\|_F}{\sqrt{n}\|\mathbf{\Delta}_J\|_F} \ge \frac{\kappa}{2},$$

which completes the proof.

E.9. Lemma 8 and its proof.

LEMMA 8. *Assume that Condition 6 and the first part of Condition 4 hold, $q \le p$, $\log p = o(n^{1/3})$, and $\lambda = c_3\sqrt{(\log p)/(nq)}$ with $c_3 > 0$ some large enough constant. Then with probability at least $1 - O(p^{-c})$ for some positive constant c, it holds that*

$$\frac{1}{nq}\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \le \frac{\lambda}{2}.$$

Proof. An application of the union bound leads to

$$P(\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \ge nq\lambda/2) \le \sum_{j=1}^{\widetilde{p}} P\left\{\sum_{k=1}^{q}(\widetilde{\mathbf{X}}^T\mathbf{W})_{jk}^2 \ge (nq\lambda/2)^2\right\} \tag{A.46}$$

for any $\lambda \ge 0$, where $(\widetilde{\mathbf{X}}^T\mathbf{W})_{jk}$ is the (j,k)-entry of $\widetilde{\mathbf{X}}^T\mathbf{W}$. The key ingredient of the proof is to bound $P\{\sum_{k=1}^{q}(\widetilde{\mathbf{X}}^T\mathbf{W})_{jk}^2 \ge (nq\lambda/2)^2\}$. Define $T_{jk,1} = \sum_{i=1}^{n}\widetilde{\mathbf{X}}_{ij}\mathbf{W}_{ik}I(|\widetilde{\mathbf{X}}_{ij}| \le L)$ and $T_{jk,2} = \sum_{i=1}^{n}\widetilde{\mathbf{X}}_{ij}\mathbf{W}_{ik}I(|\widetilde{\mathbf{X}}_{ij}| > L)$, where $L > 0$ will be specified later. Since $(\widetilde{\mathbf{X}}^T\mathbf{W})_{jk} = \sum_{i=1}^{n}\widetilde{\mathbf{X}}_{ij}\mathbf{W}_{ik} =$

$T_{jk,1} + T_{jk,2}$, we deduce

$$\begin{aligned} P\left\{\sum_{k=1}^{q}(\widetilde{\mathbf{X}}^T\mathbf{W})_{jk}^2 \geq (nq\lambda/2)^2\right\} &\leq P\left\{\sum_{k=1}^{q} T_{jk,1}^2 \geq (nq\lambda/4)^2\right\} \\ &\quad + P\left\{\sum_{k=1}^{q} T_{jk,2}^2 \geq (nq\lambda/4)^2\right\} \\ &\leq \sum_{k=1}^{q} P\left\{|T_{jk,1}| \geq \sqrt{q}n\lambda/4\right\} + P\left\{\sum_{k=1}^{q} T_{jk,2}^2 \geq (nq\lambda/4)^2\right\}. \end{aligned} \tag{A.47}$$

We will deal with the two terms on the right hand side above separately.

We first bound $P\left\{|T_{jk,1}| \geq \sqrt{q}n\lambda/4\right\}$. By the first part of Condition 4, there exist some positive constants a_1 and b_1 such that

$$P(|\mathbf{v}^T\mathbf{x}_i| > t) \leq a_1 \exp(-b_1 t^2)$$

for any $\|\mathbf{v}\|_2 = 1$ and $t > 0$, where $\mathbf{x}_i^T = (\mathbf{X}_{i1}, \cdots, \mathbf{X}_{ip})$ is the ith row of the main effect design matrix $\mathbf{X}$. Then choosing $\mathbf{v}$ as a unit vector with the jth component being 1 gives

$$P(|\mathbf{X}_{ij}| > t) \leq a_1 \exp(-b_1 t^2)$$

for any $1 \leq i \leq n$, $1 \leq j \leq p$, and $t > 0$. Thus it follows from Lemma 4 that

$$E(\mathbf{X}_{ij}^2) \leq 2(1 + a_1)/b_1 \quad \text{and} \quad E(\mathbf{X}_{ij}^4) \leq 8(1 + a_1)/b_1^2$$

for all i and j.

Note that $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{ij}$ for $1 \leq j \leq p$ and $\widetilde{\mathbf{X}}_{ij} = \mathbf{X}_{i\ell}\mathbf{X}_{i\ell'}$ with $1 \leq \ell < \ell' \leq p$ for $p+1 \leq j \leq \widetilde{p}$. Thus $E(\widetilde{\mathbf{X}}_{jk}^2)$ are uniformly bounded from above by some positive constant $\widetilde{C}_1$. Similarly, by Condition 6 and Lemma 5 there exist some positive constants a_2 and b_2 such that $E(|\mathbf{W}_{ik}|^m) \leq a_2 b_2^m m!$ for any nonnegative integer m and indices i and k. Since $\widetilde{\mathbf{X}}_{ij}$ is independent of $\mathbf{W}_{ik}$, we have

$$\begin{aligned} E\left[|\widetilde{\mathbf{X}}_{ij}\mathbf{W}_{ik} I(|\widetilde{\mathbf{X}}_{ij}| \leq L)|^m\right] &\leq L^{m-2} E(\widetilde{\mathbf{X}}_{ij}^2) E(|\mathbf{W}_{ik}|^m) \\ &\leq m!(Lb_2)^{m-2}(2a_2 b_2^2 \widetilde{C}_1)/2 \end{aligned}$$

for each integer $m \geq 2$. In view of $T_{jk,1} = \sum_{i=1}^{n} \widetilde{\mathbf{X}}_{ij}\mathbf{W}_{ik} I(|\widetilde{\mathbf{X}}_{ij}| \leq L)$, applying Bernstein's inequality (Lemma 2.2.11 of [33]) yields

$$P\left\{|T_{jk,1}| \geq \sqrt{q}n\lambda/4\right\} \leq 2\exp\left(-\frac{qn\lambda^2}{64a_2 b_2^2 \widetilde{C}_1 + 8b_2 L\sqrt{q}\lambda}\right). \tag{A.48}$$

We next bound $P\left\{\sum_{k=1}^{q} T_{jk,2}^2 \geq (nq\lambda/4)^2\right\}$. By the definition of $T_{jk,2}$, it is seen that for each j, such an event satisfies

$$\left\{\sum_{k=1}^{q} T_{jk,2}^2 \geq (nq\lambda/4)^2\right\} \subset \left\{|\widetilde{\mathbf{X}}_{ij}| > L \text{ for some } 1 \leq i \leq n\right\}.$$

Thus using the union bound, we obtain

$$P\left\{\sum_{k=1}^{q} T_{jk,2}^2 \geq (nq\lambda/4)^2\right\} \leq \sum_{i=1}^{n} P\{|\widetilde{\mathbf{X}}_{ij}| > L\}.$$

Combining this inequality with (A.46)–(A.48) gives

$$\begin{aligned}
(\text{A.49}) \quad P(\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \geq nq\lambda/2) &\leq 2\widetilde{p}q \exp\left(-\frac{qn\lambda^2}{64a_2b_2^2\widetilde{C}_1 + 8b_2L\sqrt{q}\lambda}\right) \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{p} P\{|\mathbf{X}_{ij}| > L\} + \sum_{i=1}^{n}\sum_{1\leq\ell<\ell'\leq p}^{p} P\{|\mathbf{X}_{i\ell}\mathbf{X}_{i\ell'}| > L\} \\
&\leq qp^2 \exp\left(-\frac{qn\lambda^2}{64a_2b_2^2\widetilde{C}_1 + 8b_2L\sqrt{q}\lambda}\right) + a_1np\exp(-b_1L^2) \\
&\quad + a_1np^2\exp(-b_1L).
\end{aligned}$$

Note that $\lambda = c_3\sqrt{(\log p)/(nq)}$ with some large enough positive constant c_3. Therefore, setting $L = \widetilde{C}_2\sqrt{n/(\log p)}$ for some large positive constant $\widetilde{C}_2$ ensures that there exists some positive constant c_4 such that

$$(\text{A.50}) \qquad P(\|\widetilde{\mathbf{X}}^T\mathbf{W}\|_{2,\infty} \geq nq\lambda/2) \leq O(p^{-c}),$$

where we have used the assupmtion that $q \leq p$ and $\log p = o(n^{1/3})$. This concludes the proof.

E.10. Additional lemmas.

LEMMA 9 (Hoeffding's inequality). *Let X be a real-valued random variable with $E(X) = 0$. If $P(a \leq X \leq b) = 1$ for some $a, b \in \mathbb{R}$, then $E[\exp(tX)] \leq \exp[t^2(b-a)^2/8]$ for any $t > 0$.*

LEMMA 10 (Lemma B.4 of Hao and Zhang [16]). *Let $Z_1, \cdots, Z_n$ be independent random variables with zero mean and $E[\exp(T_0|Z_i|^\alpha)] \leq A_0$ for*

constants $T_0, A_0 > 0$ *and* $0 < \alpha \leq 1$. *Then there exist some constants* $\widetilde{C}_3, \widetilde{C}_4 > 0$ *such that*

$$P\Big(\big|n^{-1}\sum_{i=1}^{n} Z_i\big| > \epsilon\Big) \leq \widetilde{C}_3 \exp(-\widetilde{C}_4 n^{\alpha}\epsilon^2)$$

for any $0 < \epsilon \leq 1$.

Department of Information Systems and Decision Sciences
Mihaylo College of Business and Economics
California State University at Fullerton
Fullerton, CA 92831
USA
E-mail: yinfeiko@usc.edu

Department of Statistics
University of Central Florida
Orlando, FL 32816-2370
USA
E-mail: daoji.li@ucf.edu

Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
USA
E-mail: fanyingy@marshall.usc.edu
jinchilv@marshall.usc.edu