Electronic Journal of Statistics
Vol. 13 (2019) 911–984
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1534

Inference for elliptical copula multivariate response regression models

Yue Zhao

Research Centre for Operations Research and Business Statistics (ORSTAT)
Katholieke Universiteit Leuven, Naamsestraat 69
Box 3555, 3000 Leuven, Belgium
e-mail: yue.zhao@kuleuven.be

and

Christian Genest

Department of Mathematics and Statistics, McGill University
805, rue Sherbrooke ouest, Montréal (Québec) Canada H3A 0B9
e-mail: Christian.Genest@mcgill.ca

Abstract: The estimation of the coefficient matrix in a multivariate response linear regression model is considered in situations where we can observe only strictly increasing transformations of the continuous responses and covariates. It is further assumed that the joint dependence between all the observed variables is characterized by an elliptical copula. Penalized estimators of the coefficient matrix are obtained in a high-dimensional setting by assuming that the coefficient matrix is either element-wise sparse or row-sparse, and by incorporating the precision matrix of the error, which is also assumed to be sparse. Estimation of the copula parameters is achieved by inversion of Kendall's tau. It is shown that when the true coefficient matrix is row-sparse, the estimator obtained via a group penalty outperforms the one obtained via a simple element-wise penalty. Simulation studies are used to illustrate this fact and the advantage of incorporating the precision matrix of the error when the correlation among the components of the error vector is strong. Moreover, the use of the normal-score rank correlation estimator is revisited in the context of high-dimensional Gaussian copula models. It is shown that this estimator remains as the optimal estimator of the copula correlation matrix in this setting.

MSC 2010 subject classifications: Primary 62F12, 62J02; secondary 62H20, 62H99.
Keywords and phrases: Copula, Dantzig selector, Kendall's tau, Lasso, normal-score rank correlation, regression, sparsity, U-statistic, van der Waerden correlation.

Received September 2017.

1. Introduction

Suppose that a $q \times 1$ response vector $\mathbf{Y}$ has been observed on a random sample of subjects and that we wish to determine whether its behavior is influenced by explanatory variables forming a $p \times 1$ vector $\mathbf{X}$ measured on the same individuals.

A standard approach to this problem consists of assuming that there exist linear relationships between the components of $\mathbf{Y}$ and those of $\mathbf{X}$, viz.

$$\mathbf{Y}^\top = \mathbf{X}^\top \mathbf{B}^* + \boldsymbol{\varepsilon}^\top, \tag{1.1}$$

where $\boldsymbol{\varepsilon}$ is a $q \times 1$ vector of errors and $\mathbf{B}^*$ is a $p \times q$ matrix of coefficients to be selected wisely. This problem has been intensely investigated, even in the recently emerging context where the dimensions of $\mathbf{X}$ and $\mathbf{Y}$ are larger than the sample size. In the latter case, various sparsity conditions on $\mathbf{B}^*$ are typically imposed in order to ensure that its estimation is feasible, reliable and efficient.

Model (1.1) is not always realistic in practice. However, it will here be shown that it remains possible to estimate $\mathbf{B}^*$ in a much broader class of models in which it is merely assumed that strictly increasing transformations of the components of $\mathbf{X}$ and $\mathbf{Y}$ (but not necessarily the original $\mathbf{X}$ and $\mathbf{Y}$ themselves) are linked by a multivariate linear regression model. To be more precise, suppose that there exist fixed but unknown functions $f_1, \ldots, f_p$ and $g_1, \ldots, g_q$ that are strictly increasing on $\mathbb{R}$ and such that

$$g(\mathbf{Y})^\top = f(\mathbf{X})^\top \mathbf{B}^* + \boldsymbol{\varepsilon}^\top, \tag{1.2}$$

where $f(X_1, \ldots, X_p) = (f_1(X_1), \ldots, f_r(X_p))^\top$ and similarly, $g(Y_1, \ldots, Y_q) = (g_1(Y_1), \ldots, g_q(Y_q))^\top$.

The estimation of $\mathbf{B}^*$ in Model (1.2) has recently been considered by Cai and Zhang [4] in the special case where $q = 1$ and the vector $(f(\mathbf{X}), g(\mathbf{Y}))$ is jointly normal and scaled in such a way that its components have unit variance. These authors provide a rate-optimal estimation procedure for the vector $\mathbf{B}^*$ which is adaptive to the unknown marginal transformations. They also use it to investigate how American crime statistics are connected to socio-economic and law enforcement data.

This paper extends the results of Cai and Zhang [4] in two ways. First, by allowing $(f(\mathbf{X}), g(\mathbf{Y}))$ to have an elliptically contoured distribution [5], we cover cases in which the variables exhibit greater tail dependence than if they were jointly normal. The multivariate Student t distribution is an example. Second, we address issues pertaining to the estimation of $\mathbf{B}^*$ that arise only when the response is of dimension $q > 1$ in Model (1.2). In this context, it is generally advisable to take into account the overall structure of the matrix $\mathbf{B}^*$ to reduce the dimension of the problem. This is especially important when p and q are comparable to, or larger than, the sample size n, in which case traditional estimators, e.g., those based on the least squares principle, are either unfeasible or perform poorly.

In the framework of Model (1.2) with elliptical dependence structure and arbitrary dimension $q > 1$, we consider the estimation of $\mathbf{B}^*$ under two conditions: row-sparsity, in which most rows of $\mathbf{B}^*$ are assumed to be zero vectors, and element-wise sparsity, where most entries of $\mathbf{B}^*$ are zero but not according to any specific pattern. Neither condition admits a univariate analog unless $\operatorname{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$ is diagonal. It is here shown that the estimation of a row- or element-wise sparse matrix $\mathbf{B}^*$ can be achieved within the broad framework of Model (1.2) without assuming that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$ is diagonal.

In the limited context of Model (1.1), Yuan and Lin [44] showed that improved estimation of a row-sparse matrix $\mathbf{B}^*$ can be achieved through a group Lasso penalty, with different groups corresponding to the different rows of $\mathbf{B}^*$. The advantage of group sparsity over simpler element-wise sparsity conditions on $\mathbf{B}^*$ is further documented in [16, 29, 32]. However, these studies generally assume that the components of $\boldsymbol{\varepsilon}$ in (1.1) are uncorrelated. To avoid this restriction, we proceed as in [35, 43], where this assumption is relaxed by requiring only that the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}^{-1}$ is sparse in Model (1.1). Both $\mathbf{B}^*$ and $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ can then be estimated using a penalized least squares with Lasso-type penalties on the two matrices. As shown numerically by Rothman et al. [35], the estimation accuracy for $\mathbf{B}^*$ is improved by incorporating an estimate of $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ when the components of the error vector $\boldsymbol{\varepsilon}$ are strongly correlated. The resulting optimization problem is only convex in $\mathbf{B}^*$ for fixed $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$, and vice versa. We face a similar issue in Model (1.2) and instead of relying on an iterative procedure to estimate the two matrices, we mimic [35] by adopting a one-step method in which an improved estimation of $\mathbf{B}^*$ is obtained after estimating $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ once.

Our main result, stated more precisely in Sections 5.3.1 and 5.3.2, is that under the row-sparsity assumption, the group penalty approach leads to a better estimation of $\mathbf{B}^*$ than the element-wise penalty approach, as already reported by Lounici et al. [29] under Model (1.1). This conclusion could not be reached by simply aggregating the univariate-response estimator considered in [4] over the multiple responses. It is remarkable in that it holds even though the estimation is based on observations from $(\mathbf{X}, \mathbf{Y})$ rather than $(f(\mathbf{X}), g(\mathbf{Y}))$. We also show through simulation the benefit of incorporating $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ when the correlation among the components of the error vector is strong.

Another complication that arises in the broader context we consider is that the raw estimator of the design matrix may fail to be positive semidefinite. The naive formulation of the Lasso program is then no longer convex. To address this issue, we can either modify the Lasso program, as motivated in [6], or adopt a Dantzig selector program. In general, the former is computationally more efficient, while the latter yields faster convergence rates under milder conditions, as we will prove. Consequently, even when $q = 1$, our approach should be preferable to that of Cai and Zhang [4]; see Section 3.3.

1.1. Plan of the paper

The model is further discussed in Section 2. The estimator of the coefficient matrix $\mathbf{B}^*$ is then developed in three stages described in Sections 3–5 as follows.

In Section 3, we consider a column-by-column estimator $\widetilde{\mathbf{B}}$ of $\mathbf{B}^*$ which does not use any information about the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$. As shown in Section 3.1, its columns are solutions to the Lasso program (3.2). An alternative approach via the Dantzig selector is described in Section 3.2.

In Section 4, the preliminary estimator $\widetilde{\mathbf{B}}$ is used to obtain an estimator $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ of the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$. The estimator $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is the solution to the graphical

Lasso algorithm (4.2), though any algorithm, such as CLIME [3] or the D-trace loss [46], yielding comparable performance could also be used.

In Section 5, an improved estimation of the coefficient matrix $\mathbf{B}^*$ is obtained by incorporating the estimator $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$. In addition to an element-wise sparsity structure, we consider a row-sparse model for $\mathbf{B}^*$ and to this end, we impose a group penalty on the coefficient matrix for its second estimation. For the element-wise sparsity case, the final estimator $\widehat{\mathbf{B}}$ of $\mathbf{B}^*$ is given in (5.4); for the row-sparse case, the final estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$ via the group Lasso approach appears in (5.7), while (5.14) gives the final estimator $\widehat{\mathbf{B}}_{\mathrm{D,G}}$ via the group Dantzig selector; see, e.g., [21, 26].

Section 6 reports the results of a modest simulation study comparing the estimators of $\mathbf{B}^*$ produced with or without consideration of the precision matrix $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$, and with or without consideration of the possible row-sparse structure of $\mathbf{B}^*$. All proofs are grouped in Section 7 and concluding comments can be found in Section 8. Some auxiliary results are deferred to an Appendix.

As an aside, we revisit in Section F the normal-score rank correlation estimator or van der Waerden correlation matrix $\boldsymbol{\Sigma}_n$, which is known to be the optimal estimator of the copula correlation matrix of a Gaussian copula in fixed dimension. We show that $\boldsymbol{\Sigma}_n$ actually retains its optimality in high-dimensional Gaussian copula models. Therefore, efficiency gains could be made by resorting to this estimator in many high-dimensional Gaussian copula modeling contexts where Kendall's tau and Spearman's rho are currently predominant. As this contribution concerns a different initial estimator of the copula component rather than the subsequent regression setup and the estimator of $\mathbf{B}^*$, its mostly self-contained presentation can be read independently from the rest of the paper.

1.2. *Notations and conventions*

Let Φ^{-1} denote the standard normal quantile function. In what follows, the Kronecker product $\otimes$ and the Hadamard product $\circ$ always take precedence over the usual matrix product. Furthermore, all functions act component-wise when applied to a vector or a matrix.

For any $r \in [0, \infty]$, $\|\cdot\|_{\ell_r}$ denotes the element-wise matrix ℓ_r norm and $\|\cdot\|_r$ is the matrix r-norm, i.e., for $\mathbf{M} \in \mathbb{R}^{a\times b}$, $\|\mathbf{M}\|_r = \max_{\mathbf{x}\in\mathbb{R}^b, \|\mathbf{x}\|_{\ell_r}\leq 1} \|\mathbf{M}\mathbf{x}\|_{\ell_r}$. In particular $\|\cdot\|_1$ is the maximum column sum and $\|\cdot\|_\infty$ is the maximum row sum. We also use $\|\cdot\|_{\mathrm{op}} \equiv \|\cdot\|_2$ to denote the operator norm. If $\mathbf{M}$ is symmetric, we write $\mathbf{M} \succeq \mathbf{0}$ if it is positive semidefinite, in which case we also write $\lambda_{\max}(\mathbf{M})$ for its largest eigenvalue (which coincides with its operator norm), $\lambda_{\min}(\mathbf{M})$ for its smallest eigenvalue, and $\mathcal{C}(\mathbf{M}) = \lambda_{\max}(\mathbf{M})/\lambda_{\min}(\mathbf{M})$ for its condition number.

Let $\mathbb{Q}$ be a generic Euclidean space, e.g., $\mathbb{Q} = \mathbb{R}^{p\times q}$ or $\mathbb{Q} = \mathbb{R}^q$. For conformal matrices or vectors $\mathbf{M}$ and $\mathbf{N}$ in $\mathbb{Q}$, we define the inner product $\langle\cdot,\cdot\rangle$ as $\langle\mathbf{M},\mathbf{N}\rangle = \mathrm{tr}(\mathbf{M}^\top\mathbf{N})$; hence in particular $\|\mathbf{M}\|_{\ell_2}^2 = \langle\mathbf{M},\mathbf{M}\rangle$. For a norm $\mathcal{R}$ defined on $\mathbb{Q}$, its dual norm $\mathcal{R}^*$ is given by

$$\mathcal{R}^*(\mathbf{M}) \equiv \sup_{\mathbf{N}\in\mathbb{Q}\setminus\{\mathbf{0}\}} \langle \mathbf{M}, \mathbf{N}\rangle / \mathcal{R}(\mathbf{N}).$$

For any $a \in \mathbb{N} = \{1, 2, \ldots\}$, we write $[a] = \{1, \ldots, a\}$. For readability, we will typically use $i, j \in [n]$ as sample indices, and $k \in [p]$ and $\ell \in [q]$ as coordinate indices, though sometimes $k, \ell \in [p+q]$. For a matrix $\mathbf{M}$, we use $(\mathbf{M})_{k\ell}$ to denote its (k, ℓ)th element, $(\mathbf{M})_{k\bullet}$ to denote its kth row, and $(\mathbf{M})_{\bullet\ell}$ to denote its ℓth column. We generally use S as an index set, sometimes with subscripts, e.g., $S \subset [p]$ or $S \subset [p] \times [q]$. If $\mathbf{a} \in \mathbb{R}^p$ and if $S \subset [p]$, we use $\mathbf{a}_S$ to denote the same vector as $\mathbf{a}$ but with entries at locations $[p] \setminus S$ set to zero. If $\mathbf{M} \in \mathbb{R}^{p\times q}$, and if $S \subset [p]$, we use $(\mathbf{M})_{S\bullet}$ to denote the same matrix as $\mathbf{M}$ but with entire rows at locations $[p] \setminus S$ set to zero, while if $S \subset [p] \times [q]$, we use $(\mathbf{M})_S$ to denote the same matrix as $\mathbf{M}$ but with entries at locations $[p] \times [q] \setminus S$ set to zero. By convention, the letter C with subscript denotes a universal constant that can be taken as fixed throughout the paper. Finally, we make the blanket assumption that the sample size n is even for simplicity.

2. Model setup and rank-based estimation

Suppose that Model (1.2) holds for fixed but unknown functions $f_1, \ldots, f_p$ and $g_1, \ldots, g_q$ that are strictly increasing on $\mathbb{R}$. For identification purpose, further assume that these functions and the scale of $\boldsymbol{\varepsilon}$ are chosen in such a way that all components of $f(\mathbf{X})$ and $g(\mathbf{Y})$ have mean 0 and variance 1. The matrix parameter $\mathbf{B}^*$ in Models (1.1)–(1.2) is then identifiable, as mentioned in Section 1 of [4].

The above identifiability conditions are not restrictive because Model (1.1) or (1.2) can always be modified to ensure that they hold; see Appendix E. Note also that while $\mathbf{B}^*$ depends on the identifiability conditions, the latter cancel with those imposed on the transformation functions f and g when the coefficient matrix enters into the important task of predicting the responses from a sample of $\mathbf{X}$. In that sense, the identifiability conditions are thus irrelevant and an accurate estimator of $\mathbf{B}^*$ will contribute to good forecasts, as we discuss in Appendix E and in the real-world example in Section 6.2. Thus we confine ourselves to the specific task of estimating $\mathbf{B}^*$ for the rest of the paper.

When the joint distribution of $(f(\mathbf{X})^\top, \boldsymbol{\varepsilon}^\top)^\top$ is normal with $f(\mathbf{X})$ independent of $\boldsymbol{\varepsilon}$, Model (1.2) is referred to as a Gaussian copula regression model. We here assume more generally that the joint distribution of $(f(\mathbf{X})^\top, \boldsymbol{\varepsilon}^\top)^\top$ has a non-degenerate elliptical distribution [5], with $f(\mathbf{X})$ uncorrelated with $\boldsymbol{\varepsilon}$. Then, the unobserved vector $(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top$ also has a non-degenerate elliptical distribution, i.e., there exists a vector $\boldsymbol{\mu} \in \mathbb{R}^{p+q}$, a nonnegative continuous random variable R and a $(p + q) \times (p + q)$ invertible matrix A such that $(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top = \boldsymbol{\mu} + RA\mathbf{U}$, where $\mathbf{U}$ is independent of R and uniformly distributed on the unit sphere in $\mathbb{R}^{p+q}$. The unique copula of the vector $(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top$ is then said to be elliptical. The properties of elliptical copulas are reviewed, e.g., in [13]. Members of this class are characterized by a univariate generator ψ (in one-to-one correspondence with the distribution of R) and a copula correlation matrix.

Given that f and g are strictly increasing, the vectors $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$ and $(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top$ have the same elliptical copula; see, e.g., Theorem 2.4.3 in [31]. For this reason, the model is called an elliptical copula multivariate response regression model. Under our identifiability conditions, the common copula correlation matrix $\mathbf{\Sigma}$ of $\mathbf{Z}$ and $(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top$ coincides with the covariance matrix of the latter. Accordingly $\mathbf{\Sigma}$ can then be estimated from the observed sample of $\mathbf{Z}$ by inversion of Kendall's tau, irrespective of ψ.

Let $\mathbf{Z}_1 = (\mathbf{X}_1^\top, \mathbf{Y}_1^\top)^\top, \ldots, \mathbf{Z}_n = (\mathbf{X}_n^\top, \mathbf{Y}_n^\top)^\top$ be a random sample of size $n \geq 2$ from $\mathbf{Z}$. Also let $\mathbf{Z} = (Z_1, \ldots, Z_{p+q})^\top$ and for each $i \in [n]$, set $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i(p+q)})^\top$. For arbitrary $k, \ell \in [p+q]$, the value of Kendall's tau between the kth and ℓth coordinates of $\mathbf{Z}$ is then defined [18, 20] as

$$\tau_{k\ell} = \mathrm{E}\{\mathrm{sgn}(Z_{1k} - Z_{2k})\,\mathrm{sgn}(Z_{1\ell} - Z_{2\ell})\}.$$

Let $\mathbf{T} \in \mathbb{R}^{(p+q)\times(p+q)}$ be the matrix whose (k,ℓ)th entry is $\tau_{k\ell}$. When $\mathbf{Z}$ is elliptical, we have

$$\mathbf{\Sigma} = \sin(\pi \mathbf{T}/2)\,, \tag{2.1}$$

as mentioned, e.g., in [17, 23]. Now consider the empirical analog $\widehat{\mathbf{T}}$ of $\mathbf{T}$, whose (k,ℓ)th entry $\widehat{\tau}_{k\ell}$ is the empirical version of Kendall's tau between the kth and ℓth coordinates of $\mathbf{Z}$, viz.

$$\widehat{\tau}_{k\ell} = \frac{2}{n(n-1)} \sum\sum_{1\leq i<j\leq n} \{\mathrm{sgn}(Z_{ik} - Z_{jk})\,\mathrm{sgn}(Z_{i\ell} - Z_{j\ell})\}.$$

A plug-in estimator of $\mathbf{\Sigma}$ is then given by

$$\widehat{\mathbf{\Sigma}} = \sin(\pi \widehat{\mathbf{T}}/2). \tag{2.2}$$

It has also long been known that $\widehat{\mathbf{T}}$ is a (matrix) U-statistic, whose limiting distribution is (matrix) Gaussian and centered at $\mathbf{T}$; see, e.g., [10]. Accordingly, $\widehat{\mathbf{\Sigma}}$ is a consistent estimator of

$$\begin{aligned}\mathbf{\Sigma} = \mathrm{cov}\{(f(\mathbf{X})^\top, g(\mathbf{Y})^\top)^\top\} &= \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{XX}} & \mathbf{\Sigma}_{\mathbf{XY}} \\ \mathbf{\Sigma}_{\mathbf{YX}} & \mathbf{\Sigma}_{\mathbf{YY}} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Sigma}_{\mathbf{XX}} & \mathbf{\Sigma}_{\mathbf{XX}}\mathbf{B}^* \\ \mathbf{B}^{*\top}\mathbf{\Sigma}_{\mathbf{XX}} & \mathbf{B}^{*\top}\mathbf{\Sigma}_{\mathbf{XX}}\mathbf{B}^* + \mathbf{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} \end{pmatrix}.\end{aligned} \tag{2.3}$$

Here the last equality in (2.3) follows from the joint ellipticity and the uncorrelatedness of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$.

Note that the plug-in estimator $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ based on Kendall's tau is not necessarily positive semidefinite; see, e.g., [42], as well as [8] for an early mention of this problem. This can be an issue in regularization routines involving $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ as a design matrix. When needed, we can rely on the projection $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ of $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ onto the set of positive semidefinite matrices, viz.

$$\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}} = \operatorname*{argmin}_{\mathbf{M}\in\mathbb{R}^{p\times p}:\mathbf{M}\succeq\mathbf{0}} \|\mathbf{M} - \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\|_{\ell_\infty}. \tag{2.4}$$

This projection can be computed efficiently using the algorithm described in Appendix A of [6]. It also comes at a minimal cost in terms of the $\|\cdot\|_{\ell_\infty}$ norm because as stated in Eq. (2.3) of [6], we have

$$\|\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \leq 2\,\|\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}. \tag{2.5}$$

As a final remark, note that in the special case of Gaussian copulas, $\widehat{\mathbf{\Sigma}}$ in (2.2) can be replaced by the better-performing normal-score rank correlation estimator $\mathbf{\Sigma}_n$ discussed in Section F.

3. First estimation of $\mathbf{B}^*$

The rank-based estimator $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ can be used to estimate the coefficient matrix $\mathbf{B}^*$ through (2.3). In this section we study a preliminary, column-by-column estimation of $\mathbf{B}^*$ that ignores any information about the precision matrix $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$. In Section 3.1, a Lasso program based on the design matrix $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ is considered; an alternative approach rooted in the Dantzig selector is described in Section 3.2.

For each $\ell \in [q]$, let $\boldsymbol{\beta}^*_\ell$ denote the ℓth column of $\mathbf{B}^*$ and let S_ℓ be the corresponding support set, i.e., the collection of indices corresponding to the nonzero elements of $\boldsymbol{\beta}^*_\ell$. For any $\alpha > 0$, further define the cone set corresponding to the ℓth column as

$$\mathbb{C}_\ell(\alpha) = \{\mathbf{x} \in \mathbb{R}^p : \|(\mathbf{x})_{S^\complement_\ell}\|_{\ell_1} \leq \alpha\,\|(\mathbf{x})_{S_\ell}\|_{\ell_1}\}.$$

The restricted eigenvalue (RE) of $\mathbf{\Sigma}_{\mathbf{XX}}$ over the cone set for the Lasso approach is defined as

$$\kappa_\ell = \min_{\mathbf{x}\in\mathbb{C}_\ell(3)} \mathbf{x}^\top \mathbf{\Sigma}_{\mathbf{XX}}\mathbf{x}/(2\,\|\mathbf{x}\|^2_{\ell_2}),$$

while for the Dantzig selector approach, we set

$$\kappa_{\mathrm{D},\ell} = \min_{\mathbf{x}\in\mathbb{C}_\ell(1)} \mathbf{x}^\top \mathbf{\Sigma}_{\mathbf{XX}}\mathbf{x}/(2\,\|\mathbf{x}\|^2_{\ell_2}).$$

Finally, let $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}$ with $\mathcal{R}(\cdot) = \|\cdot\|_{\ell_1}$ be the penalty function for the column-by-column estimation of $\mathbf{B}^*$; the dual of $\mathcal{R}$ is $\mathcal{R}^*(\cdot) = \|\cdot\|_{\ell_\infty}$.

3.1. The Lasso approach

For each $\ell \in [q]$, define the loss $\mathcal{L}_\ell : \mathbb{R}^p \to \mathbb{R}$ for the ℓth column of $\mathbf{B}^*$ by setting, for arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathcal{L}_\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}\boldsymbol{\beta}/2 - (\widehat{\mathbf{\Sigma}}_{\mathbf{XY}})^\top_{\bullet\ell}\boldsymbol{\beta}. \tag{3.1}$$

As mentioned in Section 2.2 of [4], $\mathcal{L}_\ell$ is motivated by the standard least squares loss function for the ℓth column of $\mathbf{B}^*$ in Model (1.1), but with the marginally transformed (yet unobserved) quantities $\sum_{i\in[n]} f(\mathbf{X}_i)f(\mathbf{X}_i)^\top/n$ and $\sum_{i\in[n]} f(\mathbf{X}_i)g(\mathbf{Y}_i)^\top/n$ replaced by the plug-in estimators $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XY}}$, respectively. The loss $\mathcal{L}_\ell$ is convex because $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ is positive semidefinite. The Lasso estimator of $\boldsymbol{\beta}^*_\ell$ is then given by

$$\widehat{\boldsymbol{\beta}}_\ell = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \{\mathcal{L}_\ell(\boldsymbol{\beta}) + \lambda_\ell \mathcal{R}(\boldsymbol{\beta})\}, \tag{3.2}$$

where λ_ℓ is a tuning parameter. Computationally, the term $\mathcal{L}_\ell(\boldsymbol{\beta})$ in (3.2) can be readily converted to an equivalent univariate response least squares form involving the vector $\boldsymbol{\beta}$. This transpires from the discussion around Eq. (2.2) in [6], or by treating $\mathcal{L}_\ell$ as a special case of the loss $\mathcal{L}$ in (5.1); see Appendix D. Hence (3.2) can be solved by various efficient algorithms for the standard Lasso.

The recovery rate of the estimator $\widehat{\boldsymbol{\beta}}_\ell$ is given in the following proposition, which is the analog of Theorem 1 in [4]. In what follows, C_1 is a universal constant specified just below (7.2) in Section 7.

Proposition 3.1. *Suppose that $\kappa_\ell > 0$ and n is sufficiently large to ensure that*

$$16\, C_1 s_\ell \sqrt{\ln(p^2)/n} \le \kappa_\ell/2. \tag{3.3}$$

Suppose that the tuning parameter λ_ℓ in (3.2) satisfies

$$\lambda_\ell \ge 2\, C_1 \big\{2\|\boldsymbol{\beta}^*_\ell\|_{\ell_1} \sqrt{\ln(p^2)/n} + \sqrt{\ln(pq)/n}\big\}. \tag{3.4}$$

Then, with probability at least $1 - 1/p^2 - 2/(pq)$, we have

$$\|\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell\|_{\ell_2} \le 6\sqrt{s_\ell}\, \lambda_\ell/\kappa_\ell, \quad \|\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell\|_{\ell_1} \le 24\, s_\ell \lambda_\ell/\kappa_\ell. \tag{3.5}$$

Remark 3.2. The event for which (3.5) holds is the intersection of the events $E_{\infty,1,n}$ and $E_{\infty,2,n}$ introduced at the start of Section 7.1. Observe that $E_{\infty,1,n} \cap E_{\infty,2,n}$ does not depend on $\ell \in [q]$. This fact will be used in the proofs of subsequent results.

Motivated by Proposition 3.1, our preliminary Lasso estimator of $\mathbf{B}^*$ is chosen to be

$$\widetilde{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_q). \tag{3.6}$$

3.2. The Dantzig selector approach

For each $\ell \in [q]$, define the loss $\mathcal{L}_{\mathrm{D},\ell} : \mathbb{R}^p \to \mathbb{R}$ for the ℓth column of $\mathbf{B}^*$ by setting, for arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \boldsymbol{\beta}/2 - (\widehat{\mathbf{\Sigma}}_{\mathbf{XY}})^\top_{\bullet\ell} \boldsymbol{\beta}.$$

Note that this expression is identical to (3.1), but with $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ replaced by $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$. In what follows, ∇ denotes the gradient operator. The Dantzig selector estimator of $\boldsymbol{\beta}^*_\ell$ is then given by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{D},\ell} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{R}(\boldsymbol{\beta}),$$

subject to

$$\mathcal{R}^*\{\nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta})\} \le \lambda_{\mathrm{D},\ell}, \tag{3.7}$$

where $\nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}) = \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\boldsymbol{\beta} - (\widehat{\mathbf{\Sigma}}_{\mathbf{XY}})_{\bullet\ell}$ and $\lambda_{\mathrm{D},\ell} > 0$ is a tuning parameter. Note that this Dantzig selector program is always convex, even when $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ is not positive semidefinite.

The following proposition specifies the recovery rate of the Dantzig selector estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{D},\ell}$.

Proposition 3.3. *Suppose that n is sufficiently large to ensure that*

$$\ln(2p^2) + 8\,s_\ell \ln(12p) \le n \tag{3.8}$$

and

$$8\,\mathcal{C}'(\mathbf{\Sigma}_{\mathbf{XX}})\sqrt{s_\ell \ln(12p)/n} + 2\,C_1^2 s_\ell \ln(p^2)/n \le \kappa_{\mathrm{D},\ell}, \tag{3.9}$$

where

$$\mathcal{C}'(\mathbf{\Sigma}_{\mathbf{XX}}) = 128(1+\sqrt{5})\pi\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}}), \tag{3.10}$$

where $\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})$ is the condition number of $\mathbf{\Sigma}_{\mathbf{XX}}$. Suppose that the tuning parameter $\lambda_{\mathrm{D},\ell}$ in (3.7) satisfies

$$\lambda_{\mathrm{D},\ell} \ge C_1\bigl\{\|\boldsymbol{\beta}_\ell^*\|_{\ell_1}\sqrt{\ln(p^2)/n} + \sqrt{\ln(pq)/n}\bigr\}. \tag{3.11}$$

Then, with probability at least $1 - 2/p^2 - 2/(pq)$, we have

$$\|\widehat{\boldsymbol{\beta}}_{\mathrm{D},\ell} - \boldsymbol{\beta}_\ell^*\|_{\ell_2} \le 4\sqrt{s_\ell}\,\lambda_{\mathrm{D},\ell}/\kappa_{\mathrm{D},\ell}, \quad \|\widehat{\boldsymbol{\beta}}_{\mathrm{D},\ell} - \boldsymbol{\beta}_\ell^*\|_{\ell_1} \le 8\,s_\ell\lambda_{\mathrm{D},\ell}/\kappa_{\mathrm{D},\ell}.$$

3.3. Discussion

The two approaches described above have their own merit. In general, the Lasso program is computationally more efficient than the Dantzig selector. When $\ln(p)/n \approx 0$, however, Proposition 3.1 imposes a more stringent upper bound on the number s_ℓ of nonzero elements in each column $\boldsymbol{\beta}_\ell^*$ of $\mathbf{B}^*$ than Proposition 3.3. This stems from the fact that $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^+$ given in Eq. (2.4) lacks some of the critical properties of U-statistics that $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ inherits from $\widehat{\mathbf{T}}$. It is thus more difficult to control the magnitude of $\mathbf{u}^\top(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^+ - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{u}$ than that of $\mathbf{u}^\top(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{u}$ uniformly over unit vectors $\mathbf{u}$. Refer to the proofs of Propositions 3.1 and 3.3 for details.

In practice, it may happen that $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ is positive semidefinite. When this occurs, the Lasso program inherits the same relaxed conditions as the Dantzig selector approach. Therefore, a natural question is: under what conditions is $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ likely to be positive semidefinite? Roughly speaking, the operator norm $\|\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}$ is on the order of $p/n + \sqrt{p/n}$ if we only keep factors that depend explicitly on the sample size n and the ambient dimension p; see, e.g., Lemma B.1 in the Appendix, or Theorem 2.2 in [41]. Thus, if we assume that the smallest eigenvalue of $\mathbf{\Sigma_{XX}}$ is on the order of unity, then the transition point for $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ from likely being positive semidefinite to unlikely is when p becomes larger than n. This transition is also verified by simulation studies in Section 6.

For computational sake, and because the theoretical properties of the Lasso program and the Dantzig selector differ mostly through the condition on the number of nonzero elements of the columns of $\mathbf{B}^*$, we take the Lasso approach as our starting point and use $\widetilde{\mathbf{B}}$ in (3.6) as our preliminary estimator of $\mathbf{B}^*$.

Finally we address the difference between our implementation and that of [4], who deal with the case $q = 1$. When $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ is not positive semidefinite, and transposing into our $q > 1$ context, Cai and Zhang [4] suggest either (i) to use the approach of [27] for non-convex M-estimation; or (ii) to project $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ onto the semidefinite cone to produce a positive semidefinite update $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{+,s}$ so as to minimize the operator norm of $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{+,s} - \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ when restricted to vectors with at most s nonzero elements. In our case, s will correspond to the number of nonzero elements of the single columns of $\mathbf{B}^*$.

Option (i) has the disadvantage of involving non-convex analysis and an extra tuning parameter that should bound from above the ℓ_1 norms of the single columns of $\mathbf{B}^*$; see, e.g., Theorem 3.1 in [27]. The desire to avoid such complications was a major motivation in [6] (see their discussion in Section 1), which we follow. Option (ii) is difficult to carry out in practice: first, we cannot realistically expect s to be known, and second, even if we do know s, no simple algorithm is available to carry out the required projection, in contrast to (2.4). Therefore, our column-by-column estimation of $\mathbf{B}^*$ is more straightforward and practical than that directly implied by [4].

4. Estimation of the precision matrix

In order to incorporate the precision matrix into a refined estimation of $\mathbf{B}^*$, we need to estimate $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$. Our starting point is an estimator of $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$. Then, to deduce an estimator of $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ from the estimator of $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$, we adopt the approach of [33, 34, 45]. From Eq. (2.3), we have

$$\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} = \mathbf{\Sigma_{YY}} - \mathbf{B}^{*\top}\mathbf{\Sigma_{XX}}\mathbf{B}^*.$$

Hence a plug-in estimate of $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$ is given, for $\widetilde{\mathbf{B}}$ as in (3.6), by

$$\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}} = \widehat{\mathbf{\Sigma}}_{\mathbf{YY}} - \widetilde{\mathbf{B}}^\top\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\widetilde{\mathbf{B}}. \tag{4.1}$$

By analogy with (11) in [33], we estimate $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ by the solution to the graphical Lasso program, viz.

$$\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} = \underset{\mathbf{\Omega}\in\mathbb{R}^{q\times q}:\mathbf{\Omega}\succeq\mathbf{0}}{\operatorname{argmin}} \{\langle\mathbf{\Omega}, \widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle - \ln\det(\mathbf{\Omega}) + \lambda_{\mathbf{\Omega}}\|\mathbf{\Omega}\|_{\ell_1,\text{off}}\}, \tag{4.2}$$

where $\lambda_{\mathbf{\Omega}}$ is some tuning parameter, and $\|\cdot\|_{\ell_1,\mathrm{off}}$ is the ℓ_1 norm of the entries of the argument (required to be a square matrix) excluding the diagonal elements. As is well-known, the graphical Lasso works on non-Gaussian input. Also note that the program (4.2) is always convex, even though $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ is not necessarily positive semidefinite. Program (4.2) can be solved using various graphical Lasso algorithms. If the selected algorithm requires the input to be positive semidefinite, we can simply project $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ onto the semidefinite cone in a manner analogous to (2.4) to produce an update $\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{\varepsilon\varepsilon}}$, and substitute $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ by $\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{\varepsilon\varepsilon}}$ in (4.2). Then $\widehat{\mathbf{\Sigma}}^{+}_{\boldsymbol{\varepsilon\varepsilon}}$ will satisfy an inequality analogous to (7.14) with at most an extra factor of 2 on the right-hand side, and inspection of the proof of Proposition 4.2 reveals that this leads to at most the same extra factor in the convergence rates in the proposition.

Following [33], we define the maximum degree or row cardinality of $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ as

$$d_{\mathrm{p}} = \max_{\ell\in[q]} \mathrm{card}[\{\ell' \in [q] \setminus \{\ell\} : (\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})_{\ell\ell'} \neq 0\}],$$

and let $\kappa_{\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}} = \|\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\|_\infty$. We further let $S = \{(\ell,\ell') \in [q]\times[q] : (\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})_{\ell\ell'} \neq 0\}$, $S^{\complement} = [q]\times[q]\setminus S$, and $\mathbf{\Gamma} = \mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} \in \mathbb{R}^{q^2}\times\mathbb{R}^{q^2}$. Following also the notation of [33], for any two subsets T and T' of $[q^2]$, let $(\mathbf{\Gamma})_{TT'}$ denote the $\mathrm{card}(T)\times\mathrm{card}(T')$ matrix with rows and columns of $\mathbf{\Gamma}$ indexed by T and T', respectively. Then, we set $\kappa_{\mathbf{\Gamma}} = \|(\mathbf{\Gamma})^{-1}_{SS}\|_\infty$. Finally, set

$$\Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) = \Big\{24 \max_{\ell\in[q]}(s_\ell\lambda_\ell/\kappa_\ell)\Big\}\Big\{2\|\mathbf{B}^*\|_1 + 24\max_{\ell\in[q]}(s_\ell\lambda_\ell/\kappa_\ell)\Big\} \\ + C_1\big\{\|\mathbf{B}^*\|_1^2\sqrt{\ln(p^2)/n} + \sqrt{\ln(q^2)/n}\big\}.$$

The recovery rate for the estimator $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is stated below and derived in Section 7 under the following standard irrepresentability condition, introduced in Assumption 1 in [33], for the graphic Lasso.

Assumption 4.1. *There exists* $\alpha \in (0,1]$ *such that* $\max_{e\in S^{\complement}} \|(\mathbf{\Gamma})_{\{e\}S}(\mathbf{\Gamma})^{-1}_{SS}\|_1 \leq 1-\alpha$.

Proposition 4.2. *Suppose that, for all* $\ell \in [q]$, $\kappa_\ell > 0$, λ_ℓ *in* (3.2) *satisfies* (3.4)*, and* n *is sufficiently large to ensure that* (3.3) *holds (so that Proposition 3.1 applies). Further assume that Assumption 4.1 is satisfied and that* n *is sufficiently large to ensure that*

$$6\,(1+8/\alpha)^2 \max(\kappa_{\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}}\kappa_{\mathbf{\Gamma}}, \kappa^3_{\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}}\kappa^2_{\mathbf{\Gamma}}) d \times \Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) \leq 1. \tag{4.3}$$

Finally, suppose that the tuning parameter $\lambda_{\mathbf{\Omega}}$ *in* (4.2) *satisfies*

$$\lambda_{\mathbf{\Omega}} = 8\,\Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})/\alpha. \tag{4.4}$$

Then, with probability at least $1-(1/p+1/q)^2$*, the estimator* $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ *satisfies*

$$\|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} \leq \{2\kappa_{\mathbf{\Gamma}}(1+8/\alpha)\}\Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}) \equiv \Delta_\infty(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \tag{4.5}$$

and

$$\|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \leq \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1 \leq d_{\mathrm{p}}\,\Delta_\infty(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \equiv \Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}). \tag{4.6}$$

Remark 4.3. When (4.6) holds, $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is positive semidefinite if n is large enough to ensure that

$$\Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \leq \lambda_{\min}(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}). \tag{4.7}$$

Remark 4.4. The event with probability at least $1 - (1/p + 1/q)^2$ in Proposition 4.2 is in fact the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$, for $E_{\infty,1,n}$, $E_{\infty,2,n}$ and $E_{\infty,3,n}$ introduced at the start of Section 7.1. This explicit representation will be used in the proofs of subsequent results.

5. Second estimation of the coefficient matrix

We now study improved estimates of the coefficient matrix $\mathbf{B}^*$ which exploit the estimator $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ from Section 4 and incorporate assumptions on the overall structure of $\mathbf{B}^*$. Preliminary notions are first reviewed in Section 5.1. An element-wise sparse model and a row-sparse model on $\mathbf{B}^*$ are then considered in Sections 5.2 and 5.3, respectively. A Lasso program based on $\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}}$ is used in Sections 5.2 and 5.3.1. Because the Lasso approach fails to reveal fully the benefit of using the group penalty under the row-sparse model, a group Dantzig selector program based on $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ is considered in Section 5.3.2.

5.1. Preliminaries

The loss functions $\mathcal{L}(\cdot;\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) : \mathbb{R}^{p\times q} \to \mathbb{R}$ used below to estimate $\mathbf{B}^*$ depend on matrices $\mathbf{S} \in \mathbb{R}^{p\times p}$, $\mathbf{S}_\times \in \mathbb{R}^{p\times q}$, and $\mathbf{\Omega} \in \mathbb{R}^{q\times q}$. For arbitrary $\mathbf{B} \in \mathbb{R}^{p\times q}$, set

$$\begin{aligned}\mathcal{L}(\mathbf{B};\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) &= \left\langle \mathbf{B}^\top\mathbf{S}\mathbf{B}/2 - \mathbf{S}_\times^\top\mathbf{B}, \mathbf{\Omega}\right\rangle \\ &= \operatorname{vec}(\mathbf{B})^\top \mathbf{\Omega}\otimes\mathbf{S}\operatorname{vec}(\mathbf{B})/2 - \operatorname{tr}(\mathbf{\Omega}\mathbf{S}_\times^\top\mathbf{B}).\end{aligned} \tag{5.1}$$

The quantities $\mathbf{S}$, $\mathbf{S}_\times$, $\mathbf{\Omega}$ in (5.1) will either be $\mathbf{\Sigma}_{\mathbf{XX}}$, $\mathbf{\Sigma}_{\mathbf{XY}}$ and $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$, respectively, or their estimates. When no ambiguity occurs, we write $\mathcal{L}(\cdot;\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega})$ as $\mathcal{L}$. The loss $\mathcal{L}$ is motivated by the $\mathbf{B}$-dependent component of the log-likelihood of Model (1.1) when the joint distribution of $(f(\mathbf{X}), g(\mathbf{Y}))$ is normal; see, e.g., Eq. (1.1) in [35], or Section 2.2 of [22, 43]. When the distribution of $(f(\mathbf{X}), g(\mathbf{Y}))$ is elliptical, $\mathcal{L}$ remains appropriate because $\mathcal{L}(\mathbf{B};\mathbf{\Sigma}_{\mathbf{XX}},\mathbf{\Sigma}_{\mathbf{XY}},\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ is minimized when $\mathbf{B} = \mathbf{B}^*$.

Following [30], define, for all $\mathbf{\Delta} \in \mathbb{Q}$,

$$\begin{aligned}\delta\mathcal{L}(\mathbf{\Delta},\mathbf{B}^*) &= \delta\mathcal{L}(\mathbf{\Delta},\mathbf{B}^*;\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) \\ &\equiv \mathcal{L}(\mathbf{B}^*+\mathbf{\Delta};\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) - \mathcal{L}(\mathbf{B}^*;\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) - \langle\nabla\mathcal{L}(\mathbf{B}^*;\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}),\mathbf{\Delta}\rangle \\ &= \left\langle \mathbf{\Delta}^\top\mathbf{S}\mathbf{\Delta},\mathbf{\Omega}\right\rangle/2 = \operatorname{vec}(\mathbf{\Delta})^\top\mathbf{\Omega}\otimes\mathbf{S}\operatorname{vec}(\mathbf{\Delta})/2,\end{aligned} \tag{5.2}$$

where $\nabla\mathcal{L}(\mathbf{B};\mathbf{S},\mathbf{S}_\times,\mathbf{\Omega}) = \mathbf{S}\mathbf{B}\mathbf{\Omega} - \mathbf{S}_\times\mathbf{\Omega}$. Then $\mathcal{L}$ is said to satisfy a restricted eigenvalue (RE) condition with constant $\kappa > 0$ over a set $\mathbb{C}$ if

$$\forall_{\mathbf{\Delta}\in\mathbb{C}} \quad \delta\mathcal{L}(\mathbf{\Delta},\mathbf{B}^*) \geq \kappa\,\|\mathbf{\Delta}\|_2^2. \tag{5.3}$$

5.2. Element-wise sparsity

A matrix $\mathbf{B}^*$ is called element-wise sparse if its support set $S \subset \mathbb{R}^{p\times q}$ is such that $s = \operatorname{card}(S) \ll p \times q$. Accordingly, to obtain an element-wise sparse estimator of $\mathbf{B}^*$, it is natural to regularize the least squares program with the penalty $\mathcal{R}(\cdot) = \|\cdot\|_{\ell_1}$, i.e., to estimate $\mathbf{B}^*$ by

$$\widehat{\mathbf{B}} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{p\times q}} \{\mathcal{L}(\mathbf{B}; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) + \lambda\mathcal{R}(\mathbf{B})\}, \tag{5.4}$$

where λ is a tuning parameter. As for (3.2), the term $\mathcal{L}(\mathbf{B}; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ in (5.4) can be readily converted to an equivalent univariate response least squares form involving the vectorized $\mathbf{B}$. Hence (5.4) can be solved by various efficient algorithms for the standard Lasso. See Appendix D for details.

The recovery rate for the estimator $\widehat{\mathbf{B}}$ is stated below and derived in Section 7 under the following assumption on the RE condition for the population loss function.

Assumption 5.1. *The loss $\mathcal{L}(\cdot; \boldsymbol{\Sigma}_{\mathbf{XX}}, \boldsymbol{\Sigma}_{\mathbf{XY}}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RE condition (5.3) with constant $\kappa > 0$ over the cone set $\mathbb{C} = \{\mathbf{M} \in \mathbb{R}^{p\times q} : \|(\mathbf{M})_{S^\complement}\|_{\ell_1} \leq 3\,\|(\mathbf{M})_S\|_{\ell_1}\}$.*

We define κ', the empirical counterpart to κ in Assumption 5.1, as

$$\kappa' = \kappa - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}\Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})/2 - 16\,C_1\{\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} + \Delta_\infty(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}s\sqrt{\ln(p^2)/n}. \tag{5.5}$$

Theorem 5.2. *Suppose that Assumption 5.1 and the assumptions of Proposition 4.2 hold. Further suppose*

(i) n is sufficiently large to ensure that, for κ' defined in (5.5), $\kappa' \geq \kappa/2$;
(ii) the tuning parameter λ in (5.4) satisfies

$$\lambda \geq 2\,C_1\big[2\{\|\mathbf{B}^*\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1 + \|\mathbf{B}^*\|_1\Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}\sqrt{\ln(p^2)/n} + \{\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1 + \Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}\sqrt{\ln(pq)/n}\big]. \tag{5.6}$$

Then, with probability at least $1 - (1/p + 1/q)^2$, the estimator $\widehat{\mathbf{B}}$ satisfies

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{\ell_2} \leq 6\sqrt{s}\,\lambda/\kappa, \quad \|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{\ell_1} \leq 24\,s\lambda/\kappa.$$

5.3. Row sparsity

A matrix $\mathbf{B}^*$ is called row-sparse if the set $S_G \subset [p]$ of indices corresponding to the nonzero rows of $\mathbf{B}^*$ is such that $s_{\mathrm{G}} = \operatorname{card}(S_{\mathrm{G}}) \ll p$. In order to obtain a row sparse estimator of $\mathbf{B}^*$, it is natural to regularize the least squares program by the penalty $\mathcal{R}_G(\cdot) = \|\cdot\|_{\ell_1,\ell_2}$, i.e., the ℓ_1 norm of the ℓ_2 norms of the rows of a matrix, given by

$$\|\mathbf{M}\|_{\ell_1,\ell_2} = \sum_{k=1}^{p} \|(\mathbf{M})_{k\bullet}\|_{\ell_2} = \sum_{k=1}^{p} \Big\{ \sum_{\ell=1}^{q} (\mathbf{M})_{k\ell}^2 \Big\}^{1/2}.$$

The dual of $\mathcal{R}_{\mathrm{G}}$ is $\mathcal{R}_{\mathrm{G}}^*(\cdot) = \|\cdot\|_{\ell_\infty,\ell_2}$, i.e., the maximum norm of the ℓ_2 norms of the rows of a matrix. In what follows, the appropriate cone set for the row-sparse model is given, for any $\alpha > 0$, by

$$\mathbb{C}_{\mathrm{G}}(\alpha) \equiv \{\mathbf{M} \in \mathbb{R}^{p\times q} : \|(\mathbf{M})_{S_{\mathrm{G}}^{\complement}\bullet}\|_{\ell_1,\ell_2} \le \alpha \, \|(\mathbf{M})_{S_{\mathrm{G}}\bullet}\|_{\ell_1,\ell_2}\},$$

i.e., for any matrix in the cone set $\mathbb{C}_{\mathrm{G}}(\alpha)$, the $\|\cdot\|_{\ell_1,\ell_2}$ norm of the rows of the matrix with indices in $[p] \setminus S_{\mathrm{G}}$ is bounded by α times the $\|\cdot\|_{\ell_1,\ell_2}$ norm of the rows of the same matrix with indices in S_{G}.

5.3.1. The group Lasso approach

A group Lasso estimate of $\mathbf{B}^*$ is given by

$$\widehat{\mathbf{B}}_{\mathrm{G}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times q}}{\operatorname{argmin}} \{\mathcal{L}(\mathbf{B}; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) + \lambda_{\mathrm{G}} \mathcal{R}_{\mathrm{G}}(\mathbf{B})\}, \tag{5.7}$$

where λ_{G} is a tuning parameter. As discussed immediately following (5.4), (5.7) can be efficiently tackled by any of the fast solvers for the standard group Lasso.

The recovery rate for $\widehat{\mathbf{B}}_{\mathrm{G}}$ is stated next and derived in Section 7 under the following assumption on the RE condition for the population loss function, which is the row-sparse analog of Assumption 5.1.

Assumption 5.3. *The loss $\mathcal{L}(\cdot; \boldsymbol{\Sigma}_{\mathbf{XX}}, \boldsymbol{\Sigma}_{\mathbf{XY}}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RE condition (5.3) with constant $\kappa_{\mathrm{G}} > 0$ over the cone set $\mathbb{C}_{\mathrm{G}}(3)$.*

We define the empirical counterpart to κ_{G} in Assumption 5.3 as

$$\kappa'_{\mathrm{G}} = \kappa_{\mathrm{G}} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})/2 \\ - 16\, C_1 \{\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\} s_{\mathrm{G}} \sqrt{\ln(p^2)/n}. \tag{5.8}$$

Theorem 5.4. *Suppose that Assumption 5.3 and the assumptions of Proposition 4.2 hold. Further suppose*

(i) n is sufficiently large to ensure that (4.7) holds and, for κ'_{G} defined in (5.8), $\kappa'_{\mathrm{G}} \ge \kappa_{\mathrm{G}}/2$;
(ii) the tuning parameter λ_{G} in (5.7) satisfies

$$\lambda_{\mathrm{G}} \ge 2\,\{\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\} \times \big\{2\, C_1\, \|\mathbf{B}^*\|_1 \sqrt{q \ln(p^2)/n} \\ + C_2 \sqrt{\mathcal{C}(\boldsymbol{\Sigma}_{\mathbf{YY}})/n} \sqrt{\ln(p^2) + \ln(5) q}\big\}, \tag{5.9}$$

where C_2 is the absolute constant that appears in Proposition 7.5. Then, with probability at least $1 - (1/p + 1/q)^2 - 1/p$, the estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$ satisfies

$$\|\widehat{\mathbf{B}}_{\mathrm{G}} - \mathbf{B}^*\|_{\ell_2} \leq 6\sqrt{s_{\mathrm{G}}}\,\lambda_{\mathrm{G}}/\kappa_{\mathrm{G}}, \quad \|\widehat{\mathbf{B}}_{\mathrm{G}} - \mathbf{B}^*\|_{\ell_1,\ell_2} \leq 24\, s_{\mathrm{G}}\lambda_{\mathrm{G}}/\kappa_{\mathrm{G}}.$$

In its current form, the group Lasso approach does not fully reveal the benefit of the row-sparse model. For instance, say $\mathbf{B}^*$ has exactly s_{G} nonzero rows, and all elements of these rows are nonzero. There are then qs_{G} nonzero elements in $\mathbf{B}^*$. In this case, if the tuning parameter λ is taken to be the lower limit of the specification (5.6), and if $\Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ and the parameters that are not explicitly dependent on the ambient dimensions p and q (e.g., $\|\mathbf{B}^*\|_1$) are all bounded, then by Theorem 5.2, with high probability (i.e., for p and q large), the element-wise Lasso program from Section 5.2 yields

$$\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_{\ell_2} \lesssim \sqrt{qs_{\mathrm{G}}} \times \sqrt{\ln(p)/n + \ln(pq)/n}. \tag{5.10}$$

By comparison, under the same conditions, but with the tuning parameter λ_{G} taken to be the lower limit of the specification (5.9), the first term in the curly bracket in (5.9) is dominant, and so by Theorem 5.4 the group Lasso program yields

$$\|\widehat{\mathbf{B}}_{\mathrm{G}} - \mathbf{B}^*\|_{\ell_2} \lesssim \sqrt{qs_{\mathrm{G}}} \times \sqrt{\ln(p)/n}. \tag{5.11}$$

Thus if $p \approx q$, the element-wise and the group Lasso programs yield the same rate, even though the more structured row-sparse model should intuitively give the latter an advantage, as we discuss now.

When the group Lasso approach is used in the traditional non-copula context with fixed design matrix and independent Gaussian errors, Corollary 4.1 in [29] states that the recovery rate for $\mathbf{B}^*$ under the ℓ_2 (or Frobenius) norm is bounded above, for arbitrary p and q, by a constant multiple of

$$\sqrt{s_{\mathrm{G}}/n} \times \sqrt{q + \ln(p)}. \tag{5.12}$$

This bound is strictly better than the rate (5.11) implied by Theorem 5.4, and up to a possible replacement of $\ln(p)$ by $\ln(p/s_{\mathrm{G}})$, this is also the minimax lower bound; see, e.g., the discussion below Theorem 6.1 in [29]. Moreover, under suitable conditions (e.g., the entries in the nonzero rows of $\mathbf{B}^*$ all have nonnegligible size) this upper bound is also strictly better than the lower bound achievable by the element-wise Lasso program; see, e.g., the discussion in Section 7 in [29].

For reasons similar to those given in Section 3.3, the bound on the recovery rate in Theorem 5.4 is suboptimal because $\|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2}$ is harder to control than $\|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2}$, as transpires from the proofs of Theorems 5.4 and 5.7. Also, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ may be positive semidefinite, though if the smallest eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{XX}}$ is on the order of unity, this is increasingly unlikely as p grows larger than n. On the event $\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\} = \{\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\}$, we have the following alternative to Theorem 5.4.

Proposition 5.5. *Assume the same conditions as in Theorem 5.4, except that the tuning parameter λ_{G} in (5.7) now satisfies*

$$\begin{aligned}\lambda_{\mathrm{G}} \geq 2\Big[& C_2 \big[\sqrt{\mathcal{C}(\mathbf{\Sigma_{XX}})}\{\|\mathbf{B}^*\mathbf{\Omega}_{\varepsilon\varepsilon}\|_{\mathrm{op}} + \|\mathbf{B}^*\|_{\mathrm{op}}\Delta_1(\mathbf{\Omega}_{\varepsilon\varepsilon})\} \\ & + \sqrt{\mathcal{C}(\mathbf{\Sigma_{YY}})}\{\|\mathbf{\Omega}_{\varepsilon\varepsilon}\|_{\mathrm{op}} + \Delta_1(\mathbf{\Omega}_{\varepsilon\varepsilon})\}\big]\sqrt{\{\ln(p^2)+\ln(5)q\}/n} \\ & + C_1^2\,\{\|\mathbf{B}^*\mathbf{\Omega}_{\varepsilon\varepsilon}\|_1 + \|\mathbf{B}^*\|_1\Delta_1(\mathbf{\Omega}_{\varepsilon\varepsilon})\}\sqrt{q}\,\ln(p^2)/(2n)\Big]. \end{aligned} \tag{5.13}$$

Then, on the intersection of the event $\{\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\}$ and another event with probability at least $1-(1/p+1/q)^2-9/p$, the estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$ satisfies

$$\|\widehat{\mathbf{B}}_{\mathrm{G}} - \mathbf{B}^*\|_{\ell_2} \leq 6\sqrt{s_{\mathrm{G}}}\,\lambda_{\mathrm{G}}/\kappa_{\mathrm{G}}, \quad \|\widehat{\mathbf{B}}_{\mathrm{G}} - \mathbf{B}^*\|_{\ell_1,\ell_2} \leq 24\, s_{\mathrm{G}}\lambda_{\mathrm{G}}/\kappa_{\mathrm{G}}.$$

As will be explained below Theorem 5.7, if the tuning parameter λ_{G} is taken to be the lower limit of the specification (5.13), whose right-hand side is precisely twice the right-hand side of (5.17), and under suitable conditions, the recovery rate for $\mathbf{B}^*$ stated in Proposition 5.5 under the ℓ_2 norm matches (5.12) and thus reflects the improvements brought about by the group Lasso program (over an element-wise Lasso program) under a row-sparse model. However, this result is not universally applicable, because the stated rates are only shown on the event $\{\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\}$, whose probability becomes very small when n is much smaller than p. To properly tackle the row-sparse model, it is thus preferable to estimate $\mathbf{B}^*$ using a group Dantzig selector program, as detailed next.

5.3.2. The group Dantzig selector approach

The group Dantzig selector estimator of $\mathbf{B}^*$ is defined as

$$\widehat{\mathbf{B}}_{\mathrm{D,G}} = \operatorname*{argmin}_{\mathbf{B}\in\mathbb{R}^{p\times q}} \mathcal{R}_{\mathrm{G}}(\mathbf{B}), \tag{5.14}$$

subject to

$$\mathcal{R}_{\mathrm{G}}^*\{\nabla\mathcal{L}(\mathbf{B};\widehat{\mathbf{\Sigma}}_{\mathbf{XX}},\widehat{\mathbf{\Sigma}}_{\mathbf{XY}},\widehat{\mathbf{\Omega}}_{\varepsilon\varepsilon})\} \leq \lambda_{\mathrm{D,G}}, \tag{5.15}$$

where $\lambda_{\mathrm{D,G}}$ is a tuning parameter. As already noted in Section 3.2, this Dantzig selector program is always convex even when $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ is not positive semidefinite.

The recovery rate for $\widehat{\mathbf{B}}_{\mathrm{D,G}}$ given next is derived in Section 7 under the following assumption on the RE condition for the population loss function, which is the Dantzig selector analog of Assumption 5.3.

Assumption 5.6. *The loss $\mathcal{L}(\cdot;\mathbf{\Sigma_{XX}},\mathbf{\Sigma_{XY}},\mathbf{\Omega}_{\varepsilon\varepsilon})$ satisfies RE condition (5.3) with constant $\kappa_{\mathrm{D,G}} > 0$ over the cone set $\mathbb{C}_{\mathrm{G}}(1)$.*

We define $\kappa'_{\mathrm{D,G}}$, the empirical counterpart to $\kappa_{\mathrm{D,G}}$ in Assumption 5.6, as

$$\begin{aligned}\kappa'_{\mathrm{D,G}} = 2\kappa_{\mathrm{D,G}} - \|\mathbf{\Sigma_{XX}}\|_{\mathrm{op}}\Delta_1(\mathbf{\Omega}_{\varepsilon\varepsilon}) - \big[& 2\,C_1^2\{s_{\mathrm{G}}\ln(p^2)/n\} \\ & + 8\,\mathcal{C}'(\mathbf{\Sigma_{XX}})\{s_{\mathrm{G}}\ln(12p)/n\}^{1/2}\big]\{\|\mathbf{\Omega}_{\varepsilon\varepsilon}\|_{\mathrm{op}} + \Delta_1(\mathbf{\Omega}_{\varepsilon\varepsilon})\}.\end{aligned} \tag{5.16}$$

Theorem 5.7. *Suppose that Assumption 5.6 and the assumptions of Proposition 4.2 hold. Further suppose*

(i) n is sufficiently large to ensure that (4.7), and (3.8) with s_ℓ replaced by s_{G}, hold, and for $\kappa'_{\mathrm{D,G}}$ defined in (5.16), $\kappa'_{\mathrm{D,G}} \geq \kappa_{\mathrm{D,G}}$;
(ii) the tuning parameter $\lambda_{\mathrm{D,G}}$ in (5.15) satisfies

$$\begin{aligned}\lambda_{\mathrm{D,G}} \geq C_2 \Big[&\sqrt{\mathcal{C}(\mathbf{\Sigma_{XX}})}\{\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \|\mathbf{B}^*\|_{\mathrm{op}}\Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\} \\ &+ \sqrt{\mathcal{C}(\mathbf{\Sigma_{YY}})}\{\|\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}\Big] \sqrt{\{\ln(p^2) + \ln(5)q\}/n} \\ &+ C_1^2\{\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1 + \|\mathbf{B}^*\|_1\Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}\sqrt{q}\ln(p^2)/(2n). \qquad (5.17)\end{aligned}$$

Then, with probability at least $1-(1/p+1/q)^2-9/p$, the estimator $\widehat{\mathbf{B}}_{\mathrm{D,G}}$ satisfies

$$\|\widehat{\mathbf{B}}_{\mathrm{D,G}} - \mathbf{B}^*\|_{\ell_2} \leq 4\sqrt{s_{\mathrm{G}}}\,\lambda_{\mathrm{D,G}}/\kappa_{\mathrm{D,G}}, \qquad (5.18)$$

and

$$\|\widehat{\mathbf{B}}_{\mathrm{D,G}} - \mathbf{B}^*\|_{\ell_1,\ell_2} \leq 8\,s_{\mathrm{G}}\lambda_{\mathrm{D,G}}/\kappa_{\mathrm{D,G}}. \qquad (5.19)$$

To compare the recovery rate given in Theorem 5.7 to (5.12) in terms of the ℓ_2 norm, assume for simplicity that all quantities not explicitly dependent on the ambient dimensions p and q are bounded. Further assume that the tuning parameter $\lambda_{\mathrm{D,G}}$ is the lower limit of the specification (5.17) and that

$$\text{(A) } \Delta_1(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \text{ remains bounded;} \quad \text{(B) } \ln(p)/\sqrt{n} \text{ remains bounded.} \qquad (5.20)$$

Condition (A) ensures that $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is a good estimate of $\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ while Condition (B) guarantees that the row cardinality p of $\mathbf{B}^*$ is not too large compared to n. The latter condition stems from the second term on the right-hand side of (5.17), which traces back to a second order term in the Taylor expansion of the sine transformation (2.2) from the Kendall's tau matrix estimate $\widehat{\mathbf{T}}_{\mathbf{XX}}$ to the copula correlation matrix estimate $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$.

Under these mild conditions, $\lambda_{\mathrm{D,G}}$ is on the order of $\sqrt{\{q + \ln(p)\}/n}$. Thus the recovery rate in (5.18), which is in terms of the ℓ_2 norm for estimating $\mathbf{B}^*$ under the row-sparse multivariate response elliptical copula regression model, matches the recovery rate (5.12), which is also the rate achievable for the group Lasso estimator in the traditional non-copula context with Gaussian errors. In addition, this rate is indeed superior to that achievable with the element-wise Lasso estimator.

6. Numerical performance

In this section we investigate the finite-sample numerical properties of the proposed estimators.

6.1. Simulation studies

In this simulation study, we focus exclusively on estimators for $\mathbf{B}^*$ based on the plug-in estimator (2.2) for $\boldsymbol{\Sigma}$ by inversion of Kendall's tau. Under the univariate response scenario considered in Section 4.1 of [4], when f and g are not the identity function, such rank-based estimators for $\mathbf{B}^*$ (i) routinely offer tenfold improvement in estimation accuracy compared to those based on sample covariance matrix for $\boldsymbol{\Sigma}$, because the sample covariance matrix is not robust under marginal transformations, and (ii) are almost as good as the oracle estimator that has full knowledge of f and g. Therefore, we omit comparisons with methods based on the sample covariance matrix.

We consider the following specifications for $\mathbf{B}^*$:

(a) element-wise sparse or row sparse;
(b) moderate dimension setting, $(p, q) = (20, 10)$, or high dimension setting, $(p, q) = (100, 10)$.

For the high-dimensional setting, we chose to expand $\mathbf{B}^*$ by enlarging p. Two competing factors are at work here. As discussed below Proposition 5.5 and Theorem 5.7, we know that under a row-sparse model for $\mathbf{B}^*$ and under appropriate conditions such as those stated in Eq. (5.20), the group Lasso estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$ achieves the rate (5.12) on the event $\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\}$ while the element-wise Lasso estimator $\widehat{\mathbf{B}}$ achieves the rate (5.10). It is easily checked from these rates that enlarging p increasingly favors the group Lasso estimator. When p becomes comparable to, or larger than, n, however, the event $\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\}$ very often fails. The extent to which the group Lasso estimator remains superior to the element-wise Lasso estimator when $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}$ no longer holds must then be assessed empirically.

The structure of $\mathbf{B}^*$ was obtained as follows. First, following [35], Toeplitz structures were imposed on $\boldsymbol{\Sigma}_{\mathbf{XX}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$ by setting

$$(\boldsymbol{\Sigma}_{\mathbf{XX}})_{k\ell} = 0.7^{|k-\ell|}, \quad (\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}})_{k\ell} = D\,\rho^{|k-\ell|}$$

with $D = 0.5$ and the value of ρ ranging from 0 to 0.9 to model the varying strength of the correlation among the components of the error vector $\boldsymbol{\varepsilon}$. Then $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}^{-1}$ is a tri-diagonal sparse matrix.

To obtain a row sparse model on $\mathbf{B}^*$, we first generated a $p \times q$ matrix $\widetilde{\mathbf{B}}^*$ so that exactly 80% of its p rows (chosen at random) were entirely filled with zeros. For the remaining 20% of the rows, within each column 75% of the $p/5$ elements (chosen at random within each column) were drawn from the uniform distribution on the set $\{-1, +1\}$ while the remaining 25% were made equal to zero.

To obtain an element-wise sparse model on $\mathbf{B}^*$ instead, we first generated a $p \times q$ matrix $\widetilde{\mathbf{B}}^*$ so that in each of its columns, 80% of its p elements (chosen at random within each column) are equal to zero; the remaining 20% of the entries of $\widetilde{\mathbf{B}}^*$ were drawn from the uniform distribution on $\{-1, +1\}$.

In both cases, each column of $\widetilde{\mathbf{B}}^*$ was normalized to ensure that the diagonal elements of $\widetilde{\mathbf{B}}^{*\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\widetilde{\mathbf{B}}^*$ equal $1 - D$. As a result, the diagonal elements of the sum $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}} + \widetilde{\mathbf{B}}^{*\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\widetilde{\mathbf{B}}^*$ are all 1 and hence this sum is indeed a correlation matrix. This normalized version of $\widetilde{\mathbf{B}}^*$ was taken to be the final coefficient matrix $\mathbf{B}^*$. Samples $(\mathbf{X}_1^\top, \mathbf{Y}_1^\top)^\top, \ldots, (\mathbf{X}_{100}^\top, \mathbf{Y}_{100}^\top)^\top$ were then generated from a multivariate normal distribution with covariance $\boldsymbol{\Sigma}$ given in (2.3). This served as the observed sample because Kendall's tau is invariant to increasing transforms of the margins and nothing else is needed to estimate $\mathbf{B}^*$.

Using this design, we were then able to compare the performance of three estimators:

(i) the element-wise Lasso estimator $\widehat{\mathbf{B}}$ with precision matrix included as described in Section 5.2;

(ii) the group Lasso estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$ with precision matrix included as described in Section 5.3.1;

(iii) as a benchmark, the element-wise Lasso estimator without precision matrix incorporated, obtained in the same way as $\widehat{\mathbf{B}}$ described in Section 5.2 but with precision matrix estimate $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ simply set to the identity matrix in $\mathbb{R}^{q\times q}$. This estimator of $\mathbf{B}^*$ seemed to perform better than the preliminary column-by-column estimator $\widetilde{\mathbf{B}}$ described in Section 3.1.

For the first two estimators, the tuning parameters $\lambda_1, \ldots, \lambda_q$ needed for the preliminary estimation in Section 3.1 were chosen column by column via 5-fold cross-validation, while the tuning parameter combination $(\lambda_{\boldsymbol{\Omega}}, \lambda)$ in Sections 4 and 5.2 (for the first estimator) or $(\lambda_{\boldsymbol{\Omega}}, \lambda_{\mathrm{G}})$ in Sections 4 and 5.3.1 (for the second estimator) were chosen jointly on a 2D-grid, also via 5-fold cross-validation.

Following [35], the performance of the three estimators was measured by

$$\mathrm{ME}(\widehat{\mathbf{B}}, \mathbf{B}) = \mathrm{tr}\{(\widehat{\mathbf{B}} - \mathbf{B})^\top\boldsymbol{\Sigma}_{\mathbf{XX}}(\widehat{\mathbf{B}} - \mathbf{B})\},$$

for any estimator $\widehat{\mathbf{B}}$. In each case, 500 replicates were drawn for each $\mathbf{B}^*$ and ρ specification. The median of $\mathrm{ME}(\widehat{\mathbf{B}}, \mathbf{B})$ based on the 500 repetitions is presented in Figure 1 for the three estimators.

The graphs in the left panels of Figure 1 correspond to the element-wise sparse model. They should favor the element-wise Lasso estimator over the group Lasso estimator; as ρ increases, they should also favor the estimator with precision matrix incorporated over the one without. The latter comment applies to the graphs in the right panels of Figure 1, which correspond to the row-sparse model. In this case, however, the group Lasso estimator should be preferred over the element-wise Lasso estimator, at least in the moderate dimension case with $(p, q) = (20, 10)$, where empirically only .04% of all the $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ generated (from both the element-wise sparse model and the row-sparse model, at all ρ values) failed to be positive semidefinite.

These are indeed the general trends featured in Figure 1, with a few anomalies. For the high-dimensional case with $(p, q) = (100, 10)$, none of the $\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}$ generated was positive semidefinite. Thus it may be a little surprising that even in this case, under the row-sparse model, the group Lasso estimator outperforms the element-wise Lasso estimator, by an even larger margin than under

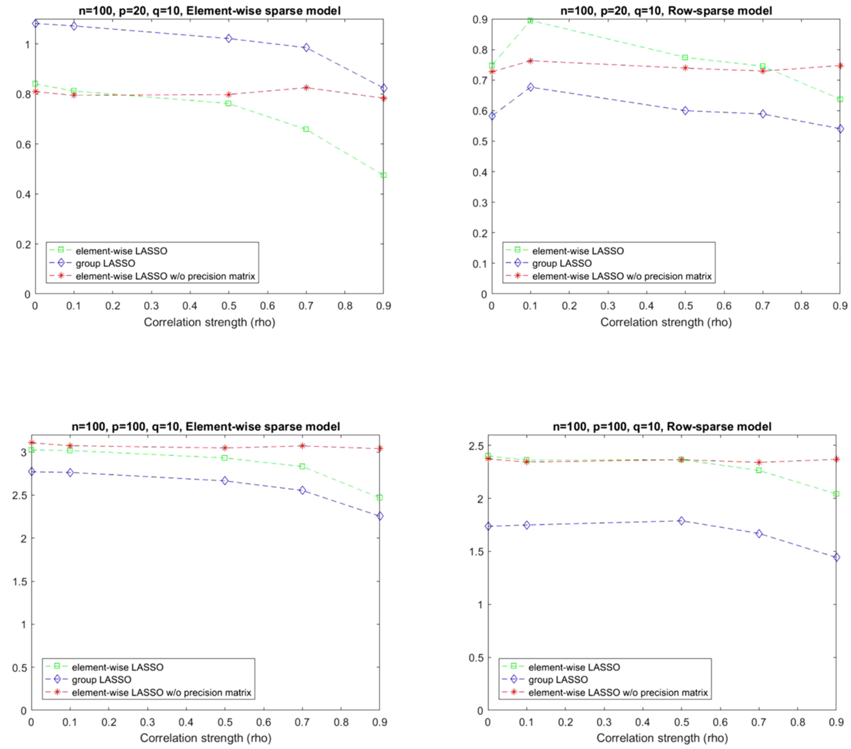

FIG 1. *Plot of* $\mathrm{ME}(\widehat{\mathbf{B}}, \mathbf{B})$ *as a function of correlation* ρ *for three estimators* $\widehat{\mathbf{B}}$*: the element-wise Lasso with precision matrix, the group Lasso with precision matrix, and the element-wise Lasso without precision matrix. Each panel corresponds to a different combination of* $(p, q) = (20, 10)$ *or* $(100, 10)$ *and the structure of* $\mathbf{B}$ *(element-wise sparse or row-sparse). Each point is based on* 500 *replicates of random samples of size* $n = 100$*.*

the moderate dimension setting. Perhaps the projection operation (2.4) does not cause much variation from $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ to $\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}$, even though (2.5) is our only theoretical guarantee. Also, it is unclear why, in this high-dimensional case and under the element-wise sparse model, the group Lasso estimator still outperforms the element-wise Lasso estimator. This should be investigated further.

6.2. Illustration

We also studied the performance of our estimators on data summarily known as the "Better Life Index," gathered for the member states of the Organization for Economic Co-operation and Development (OECD). The index "aims to involve citizens in the debate on measuring the well-being of societies, and to empower them to become more informed and engaged in the policy-making process that shapes all our lives." For each member state, the index consists of 24 socioeconomic variables. We used the data for the first $N = 34$ member states in the

2017 edition of the index, which is publicly available at https://stats.oecd.org/Index.aspx?DataSetCode=BLI

We divided the 24 variables into $q = 7$ responses and $p = 17$ explanatory variables. Three of the response variables were those labeled "Life satisfaction average score," "Quality of support network" and "Feeling safe walking alone," which are arguably subjective measures. The four other response variables were household income and wealth, as well as measures of air and water quality.

Because in practice we do not have access to the true coefficient matrix $\mathbf{B}^*$, we measured the performance of the various estimators by their predictive power. Given that the sample size is small, we compute for each member state $i \in [N]$ an empirical predictor $\widehat{\mathbf{U}}^*_{(-i)}$ using the data $\{\mathbf{x}_j : j \in [N], j \neq i\}$ from the remaining member states, and then compared the empirical predictor to the actual response $\mathbf{y}_i$. Specifically, the empirical predictor was computed as in (E.1) using

$$\widehat{f}(\mathbf{x}) = (\Phi^{-1}\{F^*_{n,1}(x_1)\}, \ldots, \Phi^{-1}\{F^*_{n,p}(x_p)\}),$$
$$\widehat{g}(\mathbf{y}) = (\Phi^{-1}\{G^*_{n,1}(y_1)\}, \ldots, \Phi^{-1}\{G^*_{n,q}(y_q)\}),$$

where Φ denotes the cdf of a standard Normal distribution, $\mathcal{N}(0,1)$, while $F^*_{n,k}$ and $G^*_{n,\ell}$ are respectively the empirical distribution functions for the kth marginal of $\mathbf{X}$ and the ℓth marginal of $\mathbf{Y}$; see (F.1). Refer to Appendix E for the justification of these choices of $\widehat{f}$ and $\widehat{g}$.

We considered seven estimators of $\mathbf{B}^*$: the element-wise Lasso estimator with precision matrix incorporated, $\widehat{\mathbf{B}}_{\mathbf{\Omega}}$; the group Lasso estimator with precision matrix incorporated $\widehat{\mathbf{B}}_{\mathrm{G},\mathbf{\Omega}}$; their counterparts $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}_{\mathrm{G}}$ without incorporating the precision matrix; the regular (i.e., without considering the copula structure) element-wise Lasso estimator $\widetilde{\mathbf{B}}$ and the regular group Lasso estimator $\widetilde{\mathbf{B}}_{\mathrm{G}}$, neither incorporating the precision matrix; and finally the ordinary least squares estimator $\widetilde{\mathbf{B}}_{\mathrm{OLS}}$. Following [11], we considered the ℓ_1 norm of the deviation $\widehat{\mathbf{U}}^*_{(-i)} - \mathbf{y}_i$ for $i \in \{1, \ldots, 34\}$.

Table 1 summarizes the performance of the empirical predictor $\widehat{\mathbf{U}}^*_{(-i)}$ for the various estimators. The results can be summarized as follows:

(a) Both the element-wise Lasso estimator $\widehat{\mathbf{B}}$ and the regular group Lasso estimator $\widetilde{\mathbf{B}}_{\mathrm{G}}$, which ignores the copula structure, perform better than the regular element-wise Lasso estimator $\widetilde{\mathbf{B}}$.

(b) The estimator $\widehat{\mathbf{B}}_{\mathrm{G}}$, which uses both the group penalty and the copula structure, thus combining the features of both $\widehat{\mathbf{B}}$ and $\widetilde{\mathbf{B}}_{\mathrm{G}}$, outperform the latter two estimators.

(c) Further incorporating precision matrix estimation to the estimators $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{B}}_{\mathrm{G}}$, which result in the estimators $\widehat{\mathbf{B}}_{\mathbf{\Omega}}$ and $\widehat{\mathbf{B}}_{\mathrm{G},\mathbf{\Omega}}$ respectively, brought mixed results.

(d) The group Lasso estimator $\widehat{\mathbf{B}}_{\mathrm{G},\mathbf{\Omega}}$ with precision matrix estimation performs somewhat better than $\widehat{\mathbf{B}}_{\mathrm{G}}$, but the element-wise Lasso estimator with precision matrix estimation $\widehat{\mathbf{B}}_{\mathbf{\Omega}}$ performs slightly worse than its counterpart $\widehat{\mathbf{B}}$.

TABLE 1
Performance of the empirical predictor $\widehat{\mathbf{U}}^*_{(-i)}$ using various estimators of $\mathbf{B}^*$ as measured by $\|\widehat{\mathbf{U}}^*_{(-i)} - \mathbf{y}_i\|_{\ell_1}$, where $i \in \{1, \ldots, 34\}$.

	$\widehat{\mathbf{B}}_{\mathbf{\Omega}}$	$\widehat{\mathbf{B}}_{\mathrm{G},\mathbf{\Omega}}$	$\widehat{\mathbf{B}}$	$\widehat{\mathbf{B}}_{\mathrm{G}}$	$\widetilde{\mathbf{B}}$	$\widetilde{\mathbf{B}}_{\mathrm{G}}$	$\widetilde{\mathbf{B}}_{\mathrm{OLS}}$
Median	40.7	35.4	40.0	37.4	53.6	46.8	60.9
80% Quantile	57.5	55.7	56.7	56.5	68.9	59.5	91.5
Sum	1417	1386	1403	1389	1838	1559	2302

Carefully checking the estimators $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ for the precision matrix reveals that the fitted $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ does not differ significantly from the identity matrix for most of the 34 empirical predictors calculated. In this illustration, therefore, the benefit of including precision matrix estimation outweighs the variation caused by the estimation for the group Lasso case, but not in the element-wise Lasso case. As a side note, personal income is consistently chosen by the estimator $\widehat{\mathbf{B}}_{\mathrm{G},\mathbf{\Omega}}$ as the variable having the most predictive power.

7. Mathematical arguments

7.1. Preliminaries

The following basic results and terminologies will be used in the proofs of the main results. First recall some basic deviation properties of Kendall's tau matrix $\widehat{\mathbf{T}}$ and the plug-in estimator $\widehat{\mathbf{\Sigma}}$. It is straightforward to show, e.g., through the equation display just above (4.28) on p. 1207 of [41], that there exist events $E_{\infty,1,n}$, $E_{\infty,2,n}$, $E_{\infty,3,n}$, with probabilities at least $1-1/p^2$, $1-2/(pq)$, $1-1/q^2$ respectively, such that

$$\|\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}}\|_{\ell_\infty} \le 2\,(C_1/\pi)\sqrt{\ln(p^2)/n}, \tag{7.1}$$
$$\|\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}}\|_{\ell_\infty} \le 2\,(C_1/\pi)\sqrt{\ln(pq)/n},$$
$$\|\widehat{\mathbf{T}}_{\mathbf{YY}} - \mathbf{T}_{\mathbf{YY}}\|_{\ell_\infty} \le 2\,(C_1/\pi)\sqrt{\ln(q^2)/n} \tag{7.2}$$

on the events $E_{\infty,1,n}$, $E_{\infty,2,n}$, $E_{\infty,3,n}$, respectively. Here C_1 is an absolute constant that can be set equal to $\sqrt{2}\pi$. Eqs. (2.1), (2.2) and the Lipschitz property of the sine function imply that

$$\|\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \le C_1\sqrt{\ln(p^2)/n}, \tag{7.3}$$
$$\|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty} \le C_1\sqrt{\ln(pq)/n}, \tag{7.4}$$
$$\|\widehat{\mathbf{\Sigma}}_{\mathbf{YY}} - \mathbf{\Sigma}_{\mathbf{YY}}\|_{\ell_\infty} \le C_1\sqrt{\ln(q^2)/n}, \tag{7.5}$$

on $E_{\infty,1,n}$, $E_{\infty,2,n}$, $E_{\infty,3,n}$, respectively. Hence, by (7.3) and (2.5), on the event $E_{\infty,1,n}$ we have

$$\|\widehat{\mathbf{\Sigma}}^+_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \le (2\,C_1)\sqrt{\ln(p^2)/n}. \tag{7.6}$$

Next, we introduce some terminology regarding the Lasso and the Dantzig selector. Given a generic Euclidean space $\mathbb{Q}$, consider a pair of subspaces $\mathcal{M} \subset \overline{\mathcal{M}}$ of $\mathbb{Q}$. We denote the orthogonal complement of $\overline{\mathcal{M}}$ by $\overline{\mathcal{M}}^{\perp}$. Following [30], we say that a norm-based penalty function $\mathcal{R} : \mathbb{Q} \to \mathbb{R}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ if

$$\forall_{\boldsymbol{\theta} \in \mathcal{M}, \boldsymbol{\gamma} \in \overline{\mathcal{M}}^{\perp}} \quad \mathcal{R}(\boldsymbol{\theta} + \boldsymbol{\gamma}) = \mathcal{R}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\gamma}).$$

Next, we define the subspace compatibility constant with respect to the pair $(\mathcal{R}, \|\cdot\|_{\ell_2})$ as

$$\Psi(\mathcal{M}) = \sup_{\boldsymbol{\theta} \in \overline{\mathcal{M}} \setminus \{\mathbf{0}\}} \mathcal{R}(\boldsymbol{\theta}) / \|\boldsymbol{\theta}\|_{\ell_2}.$$

Then, we denote the projections of $\boldsymbol{\theta} \in \mathbb{Q}$ onto $\mathcal{M}$, $\overline{\mathcal{M}}$ and $\overline{\mathcal{M}}^{\perp}$ by $\boldsymbol{\theta}_{\mathcal{M}}$, $\boldsymbol{\theta}_{\overline{\mathcal{M}}}$ and $\boldsymbol{\theta}_{\overline{\mathcal{M}}^{\perp}}$, respectively. By a cone set, we mean a subset $\mathbb{C}$ of $\mathbb{Q}$ satisfying the property that there exists $C \geq 0$ such that all $\boldsymbol{\theta} \in \mathbb{C}$ satisfies $\mathcal{R}(\boldsymbol{\theta}_{\overline{\mathcal{M}}^{\perp}}) \leq C\mathcal{R}(\boldsymbol{\theta}_{\overline{\mathcal{M}}})$.

Next, let $\mathcal{F} : \mathbb{Q} \to \mathbb{R}$ be a generic loss function and let $\boldsymbol{\beta}^*$ denote the true (but unknown) parameter value. Let also $\mathbb{C} \subset \mathbb{Q}$ be an arbitrary constraint set which will typically be a "cone set." Following [30], we define, for all $\boldsymbol{\delta} \in \mathbb{Q}$,

$$\delta\mathcal{F}(\boldsymbol{\delta}, \boldsymbol{\beta}^*) \equiv \mathcal{F}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \mathcal{F}(\boldsymbol{\beta}^*) - \langle \nabla\mathcal{F}(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle.$$

Then $\mathcal{F}$ is said to satisfy a restricted strong convexity (RSC) condition with curvature $\kappa_{\mathcal{F}}$ and tolerance function $\tau_{\mathcal{F}}$ over the set $\mathbb{C}$ if

$$\forall_{\boldsymbol{\delta} \in \mathbb{C}} \quad \delta\mathcal{F}(\boldsymbol{\delta}, \boldsymbol{\beta}^*) \geq \kappa_{\mathcal{F}} \|\boldsymbol{\delta}\|_2^2 - \tau_{\mathcal{F}}(\boldsymbol{\beta}^*). \tag{7.7}$$

If $\tau_{\mathcal{F}}(\boldsymbol{\beta}^*)$ is zero, then $\mathcal{F}$ is said to satisfy a restricted eigenvalue (RE) condition with constant $\kappa_{\mathcal{F}}$ over the set $\mathbb{C}$. Conversely, the RE condition with constant $\kappa_{\mathcal{F}}$ over a set $\mathbb{C}$ implies the RSC condition with curvature $\kappa_{\mathcal{F}}$ and tolerance function equal to zero over the same set $\mathbb{C}$. The definitions of $\delta\mathcal{L}$ and the RE condition of $\mathcal{L}$ given in Section 5.1 are special cases of these general concepts.

Finally, for the analysis of Dantzig selectors, a variant of the above RSC condition will be used. By analogy with [28], a function $\mathcal{F}$ is said to satisfy RSC-D condition with curvature $\kappa_{\mathcal{F}}$ and tolerance function $\tau_{\mathcal{F}}$ over the set $\mathbb{C}$ if

$$\forall_{\boldsymbol{\delta} \in \mathbb{C}} \quad \langle \nabla\mathcal{F}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{F}(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle \geq \kappa_{\mathcal{F}} \|\boldsymbol{\delta}\|_2^2 - \tau_{\mathcal{F}}(\boldsymbol{\beta}^*). \tag{7.8}$$

If $\mathcal{F}$ is a quadratic function, as will be the case here, the left-hand sides of (7.7) and (7.8) are both quadratic in $\boldsymbol{\delta}$ (the exact expressions differ by a factor 2), and thus the RSC and RSC-D conditions are largely identical on the same cone set $\mathbb{C}$, though they could differ further for more general loss functions. One relation between the two conditions is that, as commented on p. 565 of [28], if $\mathcal{F}$ is convex, then the RSC condition implies the RSC-D condition.

7.2. Proofs for Section 3

Proof of Proposition 3.1. The proof relies on the following result on the RSC condition.

Proposition 7.1. *On the event $E_{\infty,1,n}$ the loss function $\mathcal{L}_\ell$ satisfies RSC condition (7.7) with curvature $\kappa_\ell - 16\, C_1 s_\ell \sqrt{\ln(p^2)/n}$ and tolerance function vanishing on the cone set $\mathbb{C}_\ell(3)$.*

Proof of Proposition 7.1. We fix arbitrary $\boldsymbol{\delta} \in \mathbb{C}_\ell(3)$. It is easy to see that

$$\nabla \mathcal{L}_\ell(\boldsymbol{\beta}_\ell^*) = \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+ \boldsymbol{\beta}_\ell^* - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}})_{\bullet\ell},$$

so that in view of Definition (5.2),

$$\begin{aligned}\delta\mathcal{L}_\ell(\boldsymbol{\delta}, \boldsymbol{\beta}_\ell^*) &= \mathcal{L}_\ell(\boldsymbol{\beta}_\ell^* + \boldsymbol{\delta}) - \mathcal{L}_\ell(\boldsymbol{\beta}_\ell^*) - \langle \nabla \mathcal{L}_\ell(\boldsymbol{\beta}_\ell^*), \boldsymbol{\delta} \rangle = \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+ \boldsymbol{\delta}/2 \\ &= \boldsymbol{\delta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\delta}/2 + \boldsymbol{\delta}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+ - \boldsymbol{\Sigma}_{\mathbf{XX}}) \boldsymbol{\delta}/2.\end{aligned}$$

It then follows from two successive applications of Hölder's inequality and the definition of κ_ℓ that

$$\delta\mathcal{L}_\ell(\boldsymbol{\delta}, \boldsymbol{\beta}_\ell^*) \geq \kappa_\ell \|\boldsymbol{\delta}\|_{\ell_2}^2 - \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+ - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\boldsymbol{\delta}\|_{\ell_1}^2/2. \tag{7.9}$$

Because $\boldsymbol{\delta} \in \mathbb{C}_\ell(3)$, we also have

$$\|\boldsymbol{\delta}\|_{\ell_1} = \|\boldsymbol{\delta}_{S_\ell}\|_{\ell_1} + \|\boldsymbol{\delta}_{S_\ell^\complement}\|_{\ell_1} \leq 4\, \|\boldsymbol{\delta}_{S_\ell}\|_{\ell_1} \leq 4\sqrt{s_\ell}\, \|\boldsymbol{\delta}_{S_\ell}\|_{\ell_2}. \tag{7.10}$$

Now we focus on the event $E_{\infty,1,n}$. Plugging (7.10) into (7.9), we have

$$\begin{aligned}\delta\mathcal{L}_\ell(\boldsymbol{\delta}, \boldsymbol{\beta}_\ell^*) &\geq \kappa_\ell\, \|\boldsymbol{\delta}\|_{\ell_2}^2 - 8\, s_\ell\, \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^+ - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \|\boldsymbol{\delta}\|_{\ell_2}^2 \\ &\geq \{\kappa_\ell - 16\, C_1 s_\ell \sqrt{\ln(p^2)/n}\} \|\boldsymbol{\delta}\|_{\ell_2}^2,\end{aligned}$$

where the last step follows by (7.6), which holds on the event $E_{\infty,1,n}$. This concludes the proof. □

We wish to apply Corollary 1 in [30]. To this end, we need to (i) identify the subspaces $\mathcal{M}$ and $\overline{\mathcal{M}}$; (ii) check that the appropriate RSC condition holds; (iii) check that the tuning parameter is large enough compared to the noise level. We carry out these tasks in sequence. We focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n}$, whose probability is at least $1 - 1/p^2 - 2/(pq)$. First, we set

$$\mathcal{M} = \overline{\mathcal{M}} = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p : \forall_{k \in S_\ell^\complement}\ \beta_k = 0\}. \tag{7.11}$$

Then $\boldsymbol{\beta}_\ell^* \in \mathcal{M}$ and the penalty $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$. The subspace compatibility constant is $\Psi(\overline{\mathcal{M}}) = \sqrt{s_\ell}$.

Next, starting from Proposition 7.1, we conclude that, over the cone set $\mathbb{C}_\ell(3)$, the loss function $\mathcal{L}_\ell$ satisfies RSC condition (7.7) with tolerance function equal to zero and curvature $\kappa'_\ell = \kappa_\ell/2 > 0$, because $\kappa_\ell/2 \leq \kappa_\ell - 16\, C_1 s_\ell \sqrt{\ln(p^2)/n}$ by (3.3).

Furthermore, recall that the dual norm $\mathcal{R}^*$ of $\mathcal{R}$ is given by $\mathcal{R}^*(\cdot) = \|\cdot\|_{\ell_\infty}$. We then have

$$\begin{aligned}\mathcal{R}^*\{\nabla\mathcal{L}_\ell(\boldsymbol{\beta}^*_\ell)\} &= \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}\boldsymbol{\beta}^*_\ell - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty} \\ &= \|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\beta}^*_\ell - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty},\end{aligned}$$

and hence

$$\begin{aligned}\mathcal{R}^*\{\nabla\mathcal{L}_\ell(\boldsymbol{\beta}^*_\ell)\} &\leq \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\boldsymbol{\beta}^*_\ell\|_{\ell_1} + \|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty} \\ &\leq C_1\{2\|\boldsymbol{\beta}^*_\ell\|_{\ell_1}\sqrt{\ln(p^2)/n} + \sqrt{\ln(pq)/n}\} \leq \lambda_\ell/2.\end{aligned}$$

In the transition to the last line, we have invoked (7.4) and (7.6), and the last inequality follows by the choice of λ_ℓ in (3.4). The conclusions of the proposition then follow from Corollary 1 in [30]. □

Proof of Proposition 3.3. Set $\delta = 1/p^2$ and consider the event $E_{\mathrm{op},k,n} = E_{\mathrm{op},k,\delta,n}$ from Lemma B.1, whose probability is at least $1 - 1/p^2$. The proof relies on the following result on the RSC-D condition.

Proposition 7.2. *Suppose that n is sufficiently large to ensure that (3.8) holds. On the event $E_{\infty,1,n} \cap E_{\mathrm{op},4s_\ell,n}$, the loss function $\mathcal{L}_{\mathrm{D},\ell}$ satisfies RSC-D condition (7.8) with curvature*

$$2\,\kappa_{\mathrm{D},\ell} - \{8\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\sqrt{s_\ell \ln(12p)/n} + 2\,C_1^2 s_\ell \ln(p^2)/n\}$$

and tolerance function equal to zero over the cone set $\mathbb{C}_\ell(1)$.

Proof. We fix an arbitrary $\boldsymbol{\delta} \in \mathbb{C}_\ell(1)$. If $s_\ell = 0$, then $\boldsymbol{\delta} = \mathbf{0}$ and the conclusion of the proposition follows trivially, so we assume that $s_\ell > 0$. We have

$$\langle \nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\mathbf{XX}}\boldsymbol{\delta} + \boldsymbol{\delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\delta},$$

from which it is easy to see that

$$\langle \nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle \geq 2\,\kappa_{\mathrm{D},\ell}\|\boldsymbol{\delta}\|^2_{\ell_2} - |\boldsymbol{\delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\delta}|. \tag{7.12}$$

We fix $\delta = 1/p^2$, let $k \in \mathbb{N}$ be an arbitrary integer, and assume that (B.1) holds. We focus on the event $E_{\infty,1,n} \cap E_{\mathrm{op},k,n}$, on which (B.2) in Lemma B.1 holds with $\mathbf{u} = \boldsymbol{\delta}$ (and $\delta = 1/p^2$). We also have

$$\{\ln(2/\delta) + 2k\ln(12p)\}^{1/2} \leq 2\,\{k\ln(12p)\}^{1/2},$$

which when plugged into (B.2) (with $\mathbf{u} = \boldsymbol{\delta}$ and $\delta = 1/p^2$) yields

$$\begin{aligned}|\boldsymbol{\delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\delta}| \leq{}& [C_1^2\ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k\ln(12p)/n\}^{1/2}/k]\|\boldsymbol{\delta}\|^2_{\ell_1} \\ &+ 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\,\{k\ln(12p)/n\}^{1/2}\,\|\boldsymbol{\delta}\|^2_{\ell_2}\end{aligned}$$

$$\begin{aligned}&\le [2\,C_1^2 s_\ell \ln(p^2)/n + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{\ln(12p)/n\}^{1/2}\\ &\qquad \times (4\,s_\ell/\sqrt{k}+\sqrt{k})]\|\boldsymbol{\delta}\|_{\ell_2}^2,\end{aligned} \tag{7.13}$$

where the second inequality follows because $\boldsymbol{\delta} \in \mathbb{C}_\ell(1)$ implies $\|\boldsymbol{\delta}\|_{\ell_1}^2 \le 4\,s_\ell \|\boldsymbol{\delta}\|_{\ell_2}^2$ by a derivation similar to that of (7.10).

To balance the terms $4\,s_\ell/\sqrt{k}$ and $\sqrt{k}$ in the last line of (7.13), we choose $k = 4s_\ell$. Then (B.1) translates into (3.8), which holds by assumption, and so (7.13) also holds. Plugging (7.13) into (7.12), we then find

$$\begin{aligned}&\langle \nabla \mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla \mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle\\ &\qquad \ge [2\,\kappa_{\mathrm{D},\ell} - \{8\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\sqrt{s_\ell \ln(12p)/n} + 2\,C_1^2 s_\ell \ln(p^2)/n\}]\|\boldsymbol{\delta}\|_{\ell_2}^2,\end{aligned}$$

as claimed in the proposition. □

We wish to apply Lemma A.2. To this end, we check the three conditions at the beginning of the proof of Proposition 3.1, but with the appropriate RSC-D condition instead of the RSC condition. We focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\mathrm{op},4s_\ell,n}$ whose probability is at least $1 - 2/p^2 - 2/(pq)$.

First, we set $\mathcal{M}$ and $\overline{\mathcal{M}}$ as in (7.11) again. Then again $\boldsymbol{\beta}_\ell^* \in \mathcal{M}$, the penalty $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, and the subspace compatibility constant is $\Psi(\overline{\mathcal{M}}) = \sqrt{s_\ell}$.

Next, starting from Proposition 7.2, we conclude that, over the cone set $\mathbb{C}_\ell(1)$, the loss function $\mathcal{L}_{\mathrm{D},\ell}$ satisfies RSC-D condition (7.8) with tolerance function equal to zero and curvature $\kappa_{\mathrm{D},\ell} > 0$, because by (3.9) we have

$$\kappa_{\mathrm{D},\ell} \le 2\kappa_{\mathrm{D},\ell} - \{8\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\sqrt{s_\ell \ln(12p)/n} + 2\,C_1^2 s_\ell \ln(p^2)/n\}.$$

Furthermore, the dual norm $\mathcal{R}^*$ of $\mathcal{R}$ is given by $\mathcal{R}^*(\cdot) = \|\cdot\|_{\ell_\infty}$, and we have

$$\begin{aligned}\mathcal{R}^*\{\nabla \mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}_\ell^*)\} &= \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\boldsymbol{\beta}_\ell^* - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty}\\ &= \|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\beta}_\ell^* - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{R}^*\{\nabla \mathcal{L}_{\mathrm{D},\ell}(\boldsymbol{\beta}_\ell^*)\} &\le \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\boldsymbol{\beta}_\ell^*\|_{\ell_1} + \|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})_{\bullet\ell}\|_{\ell_\infty}\\ &\le C_1\{\|\boldsymbol{\beta}_\ell^*\|_{\ell_1}\sqrt{\ln(p^2)/n} + \sqrt{\ln(pq)/n}\} \le \lambda_{\mathrm{D},\ell},\end{aligned}$$

where the last inequality follows by the choice of $\lambda_{\mathrm{D},\ell}$ in (3.11). The conclusions of the proposition then follow from Lemma A.2. □

7.3. Proof of Proposition 4.2

Proof. We have

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} = \|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}} - \widetilde{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\widetilde{\mathbf{B}}) - (\boldsymbol{\Sigma}_{\mathbf{YY}} - \mathbf{B}^{*\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{B}^*)\|_{\ell_\infty}$$

$$\le \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}} - \boldsymbol{\Sigma}_{\mathbf{YY}}\|_{\ell_\infty} + \|\widetilde{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\widetilde{\mathbf{B}} - \mathbf{B}^{*\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{B}^*\|_{\ell_\infty}.$$

From now on we focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$ whose probability is at least $1-(1/p+1/q)^2$. Then in the last line above, the first term satisfies the bound (7.5), while for the second term,

$$\begin{aligned}
&\|\widetilde{\mathbf{B}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\widetilde{\mathbf{B}} - \mathbf{B}^{*\top}\boldsymbol{\Sigma}_{\mathbf{XX}}\mathbf{B}^*\|_{\ell_\infty} \\
&\qquad \le \|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} + \|\mathbf{B}^{*\top}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} \\
&\qquad\qquad + \|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\mathbf{B}^*\|_{\ell_\infty} + \|\mathbf{B}^{*\top}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty}.
\end{aligned}$$

We treat the terms on the right-hand side above in sequence. First,

$$\begin{aligned}
\|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} &\le \|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top\|_\infty \times \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} \\
&\le \|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top\|_\infty \times \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widetilde{\mathbf{B}}-\mathbf{B}^*\|_1 \\
&= \Big(\max_{\ell\in[q]} \|\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*_\ell\|_{\ell_1}\Big)^2,
\end{aligned}$$

and hence

$$\|(\widetilde{\mathbf{B}}-\mathbf{B}^*)^\top \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} \le \Big\{24 \max_{\ell\in[q]}(s_\ell\lambda_\ell/\kappa_\ell)\Big\}^2.$$

Here we have invoked the conclusion of Proposition 3.1 (see Remark 3.2). Next,

$$\begin{aligned}
\|\mathbf{B}^{*\top}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} &\le \|\mathbf{B}^*\|_1 \times \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}(\widetilde{\mathbf{B}}-\mathbf{B}^*)\|_{\ell_\infty} \\
&\le \|\mathbf{B}^*\|_1 \times \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widetilde{\mathbf{B}}-\mathbf{B}^*\|_1 \\
&\le 24 \max_{\ell\in[q]}(s_\ell\lambda_\ell/\kappa_\ell)\,\|\mathbf{B}^*\|_1.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\|\mathbf{B}^{*\top}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty} &\le \|\mathbf{B}^*\|_1^2 \times \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \\
&\le C_1\|\mathbf{B}^*\|_1^2\sqrt{\ln(p^2)/n}.
\end{aligned}$$

In the end, we obtain that

$$\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} \le \Delta(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}). \tag{7.14}$$

The conclusions of our proposition then follow from a slight variation of Theorem 1 in [33]. We let $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ be some generic estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$, and suppose that on some event A, the estimator $\widetilde{\widetilde{\boldsymbol{\Sigma}}}_{\boldsymbol{\varepsilon\varepsilon}}$ satisfies the error bound $\|\widetilde{\widetilde{\boldsymbol{\Sigma}}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} \le \delta$. Note that in the context and notation of Theorem 1 in [33], $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ is the observed covariance matrix for a random sample drawn directly from a distribution with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$, and $\|\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty} \le \bar{\delta}_f(n, p^\tau)$ with probability at least $1-p^{2-\tau}$.

We consider the program (4.2) with $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$ replaced by $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$. Theorem 1 in [33] states that, if the sample size is large enough so that (4.3) with $\Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})$ replaced by δ holds, and if the tuning parameter $\lambda_{\mathbf{\Omega}}$ of the program (4.2) satisfies (4.4), then on the event A the output $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ satisfies (4.5) with $\Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})$ replaced by δ, and $(\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{k\ell}$ is zero whenever $(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})_{k\ell}$ is zero. For our purpose, $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}} = \widehat{\mathbf{\Sigma}}_{\boldsymbol{\varepsilon\varepsilon}}$, the event $A = E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$, and the error bound $\delta = \Delta(\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}})$. From the above discussion, we conclude that (4.5) holds and $(\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{k\ell}$ is zero whenever $(\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})_{k\ell}$ is zero, which further imply that (4.6) hold. □

7.4. Proofs for Section 5

Proof of Theorem 5.2. The proof relies in part on the following result on the RSC condition.

Proposition 7.3. *Suppose that Assumption 5.1 and the assumptions of Proposition 4.2 hold. Then on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$ the empirical loss $\mathcal{L}(\cdot; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RSC condition (7.7) with curvature κ' introduced in (5.5) and tolerance function equal to zero over the cone set $\mathbb{C}$.*

Proof. We fix arbitrary $\mathbf{\Delta} \in \mathbb{C}$. We have

$$\delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) = \mathrm{vec}(\mathbf{\Delta})^\top \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}\, \mathrm{vec}(\mathbf{\Delta})/2,$$

and hence

$$\begin{aligned}\delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) &\\ = \mathrm{vec}(\mathbf{\Delta})^\top \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \mathbf{\Sigma}_{\mathbf{XX}}\, \mathrm{vec}(\mathbf{\Delta})/2 &\\ + \langle \mathbf{\Delta}^\top \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{\Delta}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \rangle/2 &\\ + \langle \mathbf{\Delta}^\top (\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}) \mathbf{\Delta}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} \rangle/2. &\quad (7.15)\end{aligned}$$

We treat the three terms on the right-hand side of (7.15) one by one. For the first term, by definition of $\delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \mathbf{\Sigma}_{\mathbf{XX}}, \mathbf{\Sigma}_{\mathbf{XY}}, \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ and Assumption 5.1, we have

$$\mathrm{vec}(\mathbf{\Delta})^\top \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \mathbf{\Sigma}_{\mathbf{XX}}\, \mathrm{vec}(\mathbf{\Delta})/2 = \delta\, \mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \mathbf{\Sigma}_{\mathbf{XX}}, \mathbf{\Sigma}_{\mathbf{XY}}, \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}),$$

and hence

$$\mathrm{vec}(\mathbf{\Delta})^\top \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \mathbf{\Sigma}_{\mathbf{XX}}\, \mathrm{vec}(\mathbf{\Delta})/2 \ge \kappa\, \|\mathbf{\Delta}\|^2_{\ell_2}. \qquad (7.16)$$

For the second term, we have

$$\begin{aligned}|\langle \mathbf{\Delta}^\top \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{\Delta}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \rangle| &= |\langle \mathbf{\Sigma}_{\mathbf{XX}} \mathbf{\Delta}, \mathbf{\Delta}(\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \rangle| \\ &\le \|\mathbf{\Sigma}_{\mathbf{XX}} \mathbf{\Delta}\|_{\ell_2} \times \|\mathbf{\Delta}(\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\|_{\ell_2} \\ &\le \|\mathbf{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\mathbf{\Delta}\|^2_{\ell_2}. \qquad (7.17)\end{aligned}$$

For the third term, we have

$$\begin{aligned}|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| &= |\operatorname{tr}\{\boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\}| \\ &= \Big|\sum_{\ell=1}^{q}(\boldsymbol{\Delta})^\top_{\bullet\ell}((\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\Big| \\ &\le \sum_{\ell=1}^{q}\|(\boldsymbol{\Delta})_{\bullet\ell}\|_{\ell_1}\times\|((\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_\infty},\end{aligned}$$

and hence

$$\begin{aligned}&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\ &\qquad\le \sum_{\ell=1}^{q}\|(\boldsymbol{\Delta})_{\bullet\ell}\|_{\ell_1}\times\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}\times\|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_1}. \qquad (7.18)\end{aligned}$$

Now, we bound the last factor in (7.18) as

$$\begin{aligned}\|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_1} = \|\boldsymbol{\Delta}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_1} &= \sum_{k=1}^{p}|(\boldsymbol{\Delta})_{k\bullet}(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}| \\ &\le \sum_{k=1}^{p}\|(\boldsymbol{\Delta})_{k\bullet}\|_{\ell_1}\times\|(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_\infty} = \|\boldsymbol{\Delta}\|_{\ell_1}\times\|(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_\infty},\end{aligned}$$

and therefore, continuing from (7.18), we have

$$\begin{aligned}&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\ &\qquad\le \sum_{\ell=1}^{q}\|(\boldsymbol{\Delta})_{\bullet\ell}\|_{\ell_1}\times\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}\times\|\boldsymbol{\Delta}\|_{\ell_1}\times\|(\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|_{\ell_\infty} \\ &\qquad\le \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}\times\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty}\times\|\boldsymbol{\Delta}\|^2_{\ell_1}. \qquad (7.19)\end{aligned}$$

Combining (7.15), (7.16), (7.17) and (7.19), we have

$$\begin{aligned}&\delta\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) \\ &\qquad\ge \kappa\,\|\boldsymbol{\Delta}\|^2_{\ell_2} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}\times\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}\times\|\boldsymbol{\Delta}\|^2_{\ell_2}/2 \\ &\qquad\qquad - \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}\times\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty}\times\|\boldsymbol{\Delta}\|^2_{\ell_1}/2,\end{aligned}$$

and hence, upon writing the right-hand side in a different form,

$$\begin{aligned}&\delta\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) \\ &\qquad\ge (\kappa - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}\times\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}/2)\|\boldsymbol{\Delta}\|^2_{\ell_2} \\ &\qquad\qquad - \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty}\times\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty}\times\|\boldsymbol{\Delta}\|^2_{\ell_1}/2. \qquad (7.20)\end{aligned}$$

By a derivation similar to that of (7.10), we have

$$\|\boldsymbol{\Delta}\|_{\ell_1} \le 4\,\|(\boldsymbol{\Delta})_S\|_{\ell_1} \le 4\sqrt{s}\,\|\boldsymbol{\Delta}\|_{\ell_2}. \tag{7.21}$$

Then, plugging (7.21) into (7.20), we conclude that

$$\begin{aligned}\delta\mathcal{L}(&\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon})\\ &\ge (\kappa - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - \boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_{\mathrm{op}}/2\\ &\qquad - 8\,\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_{\ell_\infty} \times s)\|\boldsymbol{\Delta}\|^2_{\ell_2}.\end{aligned} \tag{7.22}$$

Next we focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$. Then (7.6) holds, and if in addition the assumptions in Proposition 4.2 are satisfied, then Proposition 4.2 states that (4.5) and (4.6) also hold. Thus we can further bound from below the right-hand side of (7.22) as

$$\begin{aligned}&\kappa - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - \boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_{\mathrm{op}}/2 - 8\,\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_{\ell_\infty} \times s\\ &\quad \ge \kappa - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}\Delta_1(\boldsymbol{\Omega}_{\varepsilon\varepsilon})/2 - 16C_1\,\{\|\boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_{\ell_\infty} + \Delta_\infty(\boldsymbol{\Omega}_{\varepsilon\varepsilon})\}\, s\sqrt{\ln(p^2)/n}.\end{aligned}$$

This concludes the proof of Proposition 7.3. □

We wish to apply Corollary 1 in [30]. To this end, we check the three itemized conditions at the beginning of the proof of Proposition 3.1. We focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$ whose probability is at least $1 - (1/p + 1/q)^2$. First, we set

$$\mathcal{M} = \overline{\mathcal{M}} = \{\boldsymbol{\Delta} \in \mathbb{R}^{p\times q} : \forall_{(k,\ell)\in S^\complement}\ \ (\boldsymbol{\Delta})_{k\ell} = 0\}.$$

Then $\mathbf{B}^* \in \mathcal{M}$, and the penalty $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$. The subspace compatibility constant is $\Psi(\overline{\mathcal{M}}) = \sqrt{s}$. Next, by Proposition 7.3 and assumption on κ', over the cone set $\mathbb{C}$, the loss $\mathcal{L}(\cdot; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon})$ satisfies RSC condition (7.7) with tolerance function equal to zero and curvature $\kappa'' = \kappa/2 > 0$. Finally, the dual of $\mathcal{R}$ is $\mathcal{R}^*(\cdot) = \|\cdot\|_{\ell_\infty}$, and

$$\mathcal{R}^*\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon})\} = \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}\mathbf{B}^*\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_{\ell_\infty}.$$

Therefore

$$\begin{aligned}\mathcal{R}^*\{\nabla\mathcal{L}(&\mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon})\}\\ &= \|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_{\ell_\infty}\\ &\le \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\mathbf{B}^*\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_1 + \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty} \times \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon}\|_1\\ &\le \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times (\|\mathbf{B}^*\boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_1 + \|\mathbf{B}^*\|_1\|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - \boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_1)\\ &\qquad + \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty}(\|\boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_1 + \|\widehat{\boldsymbol{\Omega}}_{\varepsilon\varepsilon} - \boldsymbol{\Omega}_{\varepsilon\varepsilon}\|_1)\\ &\le \lambda/2.\end{aligned}$$

The last inequality follows by the bounds (7.6) and (7.4), Proposition 4.2, and the choice of λ in (5.6). The conclusions then follow from Corollary 1 in [30]. □

Proof of Theorem 5.4. The proof partially relies on the following result on the RSC condition.

Proposition 7.4. *Suppose that Assumption 5.3 holds, the assumptions of Proposition 4.2 hold, and n is sufficiently large to ensure that (4.7) holds. Then on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$ the loss $\mathcal{L}(\cdot; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RSC condition (7.7) with curvature κ'_{G} introduced in (5.8) and tolerance function equal to zero over the cone set $\mathbb{C}_{\mathrm{G}}(3)$.*

Proof. We fix arbitrary $\mathbf{\Delta} \in \mathbb{C}_{\mathrm{G}}(3)$, and use the same decomposition (7.15) in the proof of Proposition 7.3. The treatment for the second term in the last line of (7.15) remains the same as (7.17). For the first term, by definition of $\delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \mathbf{\Sigma}_{\mathbf{XX}}, \mathbf{\Sigma}_{\mathbf{XY}}, \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ and Assumption 5.3, we have

$$\begin{aligned}\operatorname{vec}(\mathbf{\Delta})^{\top} \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \mathbf{\Sigma}_{\mathbf{XX}} \operatorname{vec}(\mathbf{\Delta})/2 &= \delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \mathbf{\Sigma}_{\mathbf{XX}}, \mathbf{\Sigma}_{\mathbf{XY}}, \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \\ &\geq \kappa_{\mathrm{G}} \|\mathbf{\Delta}\|^2_{\ell_2}. \end{aligned} \tag{7.23}$$

For the third term, suppose that $\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is positive semidefinite. Then we can take its symmetric positive semidefinite square root $\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}}$ and use the fact that

$$\begin{aligned}|\langle \mathbf{\Delta}^{\top}(\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{\Delta}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| &= |\operatorname{tr}\{(\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}})^{\top}(\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}}\}| \\ &= \Big|\sum_{\ell=1}^{q} (\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}})^{\top}_{\bullet\ell}((\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\Big|\end{aligned}$$

to deduce that

$$\begin{aligned}&|\langle \mathbf{\Delta}^{\top}(\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{\Delta}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\ &\qquad \leq \|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \sum_{\ell=1}^{q} \|(\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}})_{\bullet\ell}\|^2_{\ell_1} \\ &\qquad \leq \|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\mathbf{\Delta}\widehat{\mathbf{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}}\|^2_{\ell_1,\ell_2} \\ &\qquad \leq \|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\mathbf{\Delta}\|^2_{\ell_1,\ell_2}. \end{aligned} \tag{7.24}$$

The second inequality follows by (B.4) in Proposition B.2. Therefore, combining (7.15), (7.23), (7.17) and (7.24) we have

$$\begin{aligned}&\delta\mathcal{L}(\mathbf{\Delta}, \mathbf{B}^*; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) \\ &\qquad \geq \kappa_{\mathrm{G}}\|\mathbf{\Delta}\|^2_{\ell_2} - \|\mathbf{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\mathbf{\Delta}\|^2_{\ell_2}/2 \\ &\qquad\qquad - \|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\mathbf{\Delta}\|^2_{\ell_1,\ell_2}/2 \\ &\qquad = (\kappa_{\mathrm{G}} - \|\mathbf{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}/2) \times \|\mathbf{\Delta}\|^2_{\ell_2} \\ &\qquad\qquad - \|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\mathbf{\Delta}\|^2_{\ell_1,\ell_2}/2. \end{aligned} \tag{7.25}$$

Because $\mathbf{\Delta} \in \mathbb{C}_{\mathrm{G}}(3)$,

$$\|\mathbf{\Delta}\|_{\ell_1,\ell_2} = \|(\mathbf{\Delta})_{S_{\mathrm{G}}\bullet}\|_{\ell_1,\ell_2} + \|(\mathbf{\Delta})_{S^{\complement}_{\mathrm{G}}\bullet}\|_{\ell_1,\ell_2}$$

$$\leq 4\,\|(\boldsymbol{\Delta})_{S_{\mathrm{G}}\bullet}\|_{\ell_1,\ell_2} \leq 4\sqrt{s_{\mathrm{G}}}\|(\boldsymbol{\Delta})_{S_{\mathrm{G}}\bullet}\|_{\ell_2} \leq 4\sqrt{s_{\mathrm{G}}}\|\boldsymbol{\Delta}\|_{\ell_2}. \tag{7.26}$$

Then, plugging (7.26) into (7.25), we conclude that

$$\begin{aligned}\delta\mathcal{L}(\boldsymbol{\Delta},\mathbf{B}^*;\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}},\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})&\\ \geq (\kappa_{\mathrm{G}} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}/2&\\ - 8\,\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} s_{\mathrm{G}}) \times \|\boldsymbol{\Delta}\|^2_{\ell_2}.&\end{aligned} \tag{7.27}$$

Now we focus on the event $E_{\infty,1,n}\cap E_{\infty,2,n}\cap E_{\infty,3,n}$. Then (7.6) holds, and if in addition the assumptions in Proposition 4.2 are satisfied, then Proposition 4.2 states that (4.6) also holds. Moreover, if in addition (4.7) holds, then $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is indeed positive semidefinite. Then (7.27) holds, and furthermore we can lower bound the right-hand side of (7.27) as

$$\begin{aligned}\kappa_{\mathrm{G}} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}/2&\\ - 8\,\|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} s_{\mathrm{G}}&\\ \geq \kappa_{\mathrm{G}} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}}\Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})/2&\\ - 16\,C_1\{\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \Delta_1(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})\}s_{\mathrm{G}}\sqrt{\ln(p^2)/n}.&\end{aligned}$$

This concludes the proof of Proposition 7.4. □

We wish to apply Corollary 1 in [30]. To this end, we check the three itemized conditions at the beginning of the proof of Proposition 3.1. We focus on the event $E_{\infty,1,n}\cap E_{\infty,2,n}\cap E_{\infty,3,n}$. First set

$$\mathcal{M} = \overline{\mathcal{M}} = \{\boldsymbol{\Delta}\in\mathbb{R}^{p\times q} : \forall_{k\in[p]\setminus S_{\mathrm{G}}^{\mathtt{C}}}\ (\boldsymbol{\Delta})_{k\bullet} = 0\}. \tag{7.28}$$

Then $\mathbf{B}^*\in\mathcal{M}$, and the penalty $\mathcal{R}_{\mathrm{G}}$ is decomposable with respect to $(\mathcal{M},\overline{\mathcal{M}}^{\perp})$. The subspace compatibility constant is $\Psi(\overline{\mathcal{M}}) = \sqrt{s_{\mathrm{G}}}$. Next, by Proposition 7.4 and assumption on κ'_{G}, over the cone set $\mathbb{C}_{\mathrm{G}}(3)$, the loss $\mathcal{L}(\cdot;\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}},\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RSC condition (7.7) with tolerance function equal to zero and curvature $\kappa''_{\mathrm{G}} = \kappa_{\mathrm{G}}/2 > 0$. Finally, we verify that

$$\begin{aligned}\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*;\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}},\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\}&\\ = \|\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}}\mathbf{B}^*\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2}&\\ = \|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}})\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2},&\end{aligned}$$

so that

$$\begin{aligned}\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*;\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}},\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}},\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\}&\\ \leq \{\|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2} + \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}\}\|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}&\\ \leq \{\sqrt{q}\,\|(\widehat{\boldsymbol{\Sigma}}^+_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty} + \|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}\}&\\ \times (\|\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}})&\end{aligned}$$

$$\leq \{\sqrt{q}\,\|\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}\|_{\ell_\infty} \times \|\mathbf{B}^*\|_1 + \|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}\} \\ \times (\|\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}). \tag{7.29}$$

The necessary bound for $\|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}$, which appears in (7.29), is given in the next proposition.

Proposition 7.5. *There exists an event $F_{1,n}$ with probability at least $1 - 1/p$ such that on the event $F_{1,n}$ we have, for some absolute constant C_2,*

$$\|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2} \leq C_2\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})}\sqrt{\{\ln(p^2) + \ln(5)q\}/n}.$$

Proof. The proof can be found in Section C.1. □

From now on we focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n} \cap F_{1,n}$, whose probability is at least $1-(1/p+1/q)^2-1/p$. Continuing from (7.29), and invoking the bound (7.6), Propositions 4.2 and 7.5, and the choice (5.9) of λ_{G}, it is easy to verify that

$$\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} \leq \lambda_{\mathrm{G}}/2.$$

The conclusions of the theorem then follow from Corollary 1 in [30]. □

Proof of Proposition 5.5. Our proof is largely similar to that of Theorem 5.4. However, instead of bounding $\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\}$, we will bound $\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\}$. Later we will focus on the event $\{\widehat{\mathbf{\Sigma}}^{+}_{\mathbf{XX}} = \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}\}$, and on this event the two bounds coincide.

Starting with a derivation similar to that in (7.29), we have

$$\begin{aligned}
&\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} \\
&\quad = \|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - (\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}})\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2} \\
&\quad\quad \leq \|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2} \\
&\quad\quad\quad + \|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}},
\end{aligned}$$

and hence

$$\begin{aligned}
&\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}, \widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} \\
&\quad \leq \|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2} \\
&\quad\quad + \|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2} \times \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \\
&\quad\quad\quad + \|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}(\|\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}).
\end{aligned} \tag{7.30}$$

In (7.30), the necessary bound for $\|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}$ has been provided in Proposition 7.5; the necessary bounds for $\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2}$ and $\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2}$ will be obtained next.

Proposition 7.6. *Let $\mathbf{D} \in \mathbb{R}^{p\times q}$ be an arbitrary non-random matrix. There exists an event $F_{\mathbf{D},n}$ with probability at least $1 - 4/p$ such that on the event $E_{\infty,1,n} \cap F_{\mathbf{D},n}$ we have*

$$\begin{aligned}&\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2}\\ &\qquad \le C_2\|\mathbf{D}\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad\qquad + C_1^2\|\mathbf{D}\|_1\sqrt{q}\ln(p^2)/(2n).\end{aligned}$$

Here C_2 is the same absolute constant as introduced in Proposition 7.5.

Proof. The proof can be found in Section C.2. □

From Proposition 7.6, where we take $\mathbf{D}$ to be $\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$ or $\mathbf{B}^*$, we conclude that there exist events $F_{2,n}$ and $F_{3,n}$, each with probability at least $1 - 4/p$, such that we have

$$\begin{aligned}&\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\ell_\infty,\ell_2}\\ &\qquad \le C_2\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad\qquad + C_1^2\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1\sqrt{q}\ \ln(p^2)/(2n),\end{aligned} \tag{7.31}$$

and

$$\begin{aligned}&\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{B}^*\|_{\ell_\infty,\ell_2}\\ &\qquad \le C_2\|\mathbf{B}^*\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad\qquad + C_1^2\|\mathbf{B}^*\|_1\sqrt{q}\ \ln(p^2)/(2n),\end{aligned} \tag{7.32}$$

respectively on $E_{\infty,1,n} \cap F_{2,n}$ and $E_{\infty,1,n} \cap F_{3,n}$.

From now on we further focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n} \cap F_{1,n} \cap F_{2,n} \cap F_{3,n}$. Starting from (7.30), (7.31), (7.32) and Proposition 7.5, we then have

$$\begin{aligned}&\mathcal{R}_{\mathrm{G}}^*\{\nabla\mathcal{L}(\mathbf{B}^*;\widehat{\mathbf{\Sigma}}_{\mathbf{XX}},\widehat{\mathbf{\Sigma}}_{\mathbf{XY}},\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\}\\ &\quad \le C_2\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad + C_1^2\|\mathbf{B}^*\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_1\sqrt{q}\ \ln(p^2)/(2n)\\ &\qquad + \{C_2\|\mathbf{B}^*\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad + C_1^2\|\mathbf{B}^*\|_1\sqrt{q}\ \ln(p^2)/(2n)\}\|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}\\ &\qquad + C_2\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n}\\ &\qquad \times(\|\mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} + \|\widehat{\mathbf{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \mathbf{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}}) \le \lambda_{\mathrm{G}}/2.\end{aligned} \tag{7.33}$$

The last step follows by Proposition 4.2 and the choice of λ_{G} in (5.13).

Next, focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n} \cap F_{1,n} \cap F_{2,n} \cap F_{3,n}$, whose probability is at least $1 - (1/p + 1/q)^2 - 9/p$, intersected with the event

$\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \succeq \mathbf{0}\} = \{\widehat{\boldsymbol{\Sigma}}^{+}_{\mathbf{XX}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\}$. The proof of Theorem 5.4 entirely goes through (on this smaller event), with the new bound $\mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}^{+}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} = \mathcal{R}^*_{\mathrm{G}}\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} \le \lambda_{\mathrm{G}}/2$ by (7.33). The conclusions of the proposition then follow. □

Proof of Theorem 5.7. The proof partially relies on the following result on the RSC-D condition.

Proposition 7.7. *Suppose that Assumption 5.6 and the assumptions of Proposition 4.2 hold, and that n is large enough to ensure that (4.7), and (3.8) with s_ℓ replaced by s_{G}, hold. Then on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n} \cap E_{\mathrm{op},4s_{\mathrm{G}},n}$, where $E_{\mathrm{op},4s_{\mathrm{G}},n}$ is the same event as $E_{\mathrm{op},k,n}$ introduced above Proposition 7.2 with $k = 4s_{\mathrm{G}}$, the empirical loss $\mathcal{L}(\cdot; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RSC-D condition (7.8) with curvature $\kappa'_{\mathrm{D,G}}$ in (5.16) and tolerance function equal to zero over the cone set $\mathbb{C}_{\mathrm{G}}(1)$.*

Proof. We have

$$\begin{aligned}\langle \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^* + \boldsymbol{\Delta}; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) &\\ - \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}), \boldsymbol{\Delta}\rangle &= \langle \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \boldsymbol{\Delta} \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}, \boldsymbol{\Delta}\rangle,\end{aligned}$$

and hence

$$\begin{aligned}&\langle \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^* + \boldsymbol{\Delta}; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) - \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}), \boldsymbol{\Delta}\rangle \\ &\qquad = \operatorname{vec}(\boldsymbol{\Delta})^\top \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \boldsymbol{\Sigma}_{\mathbf{XX}} \operatorname{vec}(\boldsymbol{\Delta}) + \langle \boldsymbol{\Delta}^\top \boldsymbol{\Sigma}_{\mathbf{XX}} \boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\rangle \\ &\qquad\qquad + \langle \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}) \boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle, \end{aligned} \tag{7.34}$$

which is identical to (7.15) apart from a factor of 2.

We fix $\boldsymbol{\Delta} \in \mathbb{C}_{\mathrm{G}}(1)$, and focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$. If $s_{\mathrm{G}} = 0$, then $\boldsymbol{\Delta} = 0$ and the conclusion of the proposition follows trivially, so we assume that $s_{\mathrm{G}} > 0$. The treatment for the second term in the last line of (7.34) remains the same as (7.17). For the first term, by definition of $\delta\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \boldsymbol{\Sigma}_{\mathbf{XX}}, \boldsymbol{\Sigma}_{\mathbf{XY}}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}})$ and Assumption 5.6, we have

$$\begin{aligned}&\operatorname{vec}(\boldsymbol{\Delta})^\top \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}} \otimes \boldsymbol{\Sigma}_{\mathbf{XX}} \operatorname{vec}(\boldsymbol{\Delta}) \\ &\qquad = 2\,\delta\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \boldsymbol{\Sigma}_{\mathbf{XX}}, \boldsymbol{\Sigma}_{\mathbf{XY}}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}) \ge 2\,\kappa_{\mathrm{D,G}} \|\boldsymbol{\Delta}\|^2_{\ell_2}. \end{aligned} \tag{7.35}$$

For the third term, because the assumptions stated in Proposition 4.2 are satisfied, (4.6) holds. Together with (4.7), this further implies that $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}$ is positive semidefinite. Hence we can take its symmetric positive semidefinite square root $\widehat{\boldsymbol{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}}$ and use

$$\begin{aligned}&|\langle \boldsymbol{\Delta}^\top (\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}) \boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\ &\qquad = |\operatorname{tr}\{(\boldsymbol{\Delta} \widehat{\boldsymbol{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}})^\top (\widehat{\boldsymbol{\Sigma}}^{+}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}}) \boldsymbol{\Delta} \widehat{\boldsymbol{\Omega}}^{1/2}_{\boldsymbol{\varepsilon\varepsilon}}\}|\end{aligned}$$

$$\leq \sum_{\ell=1}^{q} |(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}^{+} - \boldsymbol{\Sigma}_{\mathbf{XX}})(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}|. \tag{7.36}$$

Now, fix an arbitrary integer $k \in \mathbb{N}$, and assume that (B.1) holds with $\delta = 1/p^2$. In addition to the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n}$, we further focus on the event $E_{\mathrm{op},k,n}$. Then, (B.2) in Lemma B.1 applies with $\delta = 1/p^2$, and the first step of (7.13) with the substitution of $\boldsymbol{\delta}$ by $(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}$ yields

$$\begin{aligned}
&|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}| \\
&\quad \leq [C_1^2 \ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2}/k] \times \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_1}^2 \\
&\qquad + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\,\{k_\ell \ln(12p)/n\}^{1/2} \times \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_2}^2.
\end{aligned} \tag{7.37}$$

Plugging (7.37) into (7.36), we have

$$\begin{aligned}
&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\
&\quad \leq \sum_{\ell=1}^{q} \Big[C_1^2 \ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2}/k\Big] \times \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_1}^2 \\
&\qquad + \sum_{\ell=1}^{q} 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2} \times \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_2}^2.
\end{aligned}$$

Now the right-hand term can be rewritten as

$$\begin{aligned}
&\Big[C_1^2 \ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2}/k\Big] \sum_{\ell=1}^{q} \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_1}^2 \\
&\qquad + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2} \sum_{\ell=1}^{q} \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_2}^2,
\end{aligned}$$

and hence, by applying (B.4) in Proposition B.2 to the term $\sum_{\ell=1}^{q} \|(\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2})_{\bullet\ell}\|_{\ell_1}^2$, we find that it is bounded from above by

$$\begin{aligned}
&\Big[C_1^2 \ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2}/k\Big] \times \|\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2}\|_{\ell_1,\ell_2}^2 \\
&\qquad + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2} \times \|\boldsymbol{\Delta}\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}^{1/2}\|_{\ell_2}^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\
&\quad \leq [C_1^2 \ln(p^2)/(2n) + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2}/k] \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\boldsymbol{\Delta}\|_{\ell_1,\ell_2}^2 \\
&\qquad + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{k \ln(12p)/n\}^{1/2} \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\boldsymbol{\Delta}\|_{\ell_2}^2.
\end{aligned} \tag{7.38}$$

By a derivation similar to that of (7.26), now for $\boldsymbol{\Delta} \in \mathbb{C}_{\mathrm{G}}(1)$, we have

$$\|\boldsymbol{\Delta}\|_{\ell_1,\ell_2} \leq 2\sqrt{s_{\mathrm{G}}}\|\boldsymbol{\Delta}\|_{\ell_2}. \tag{7.39}$$

Then, plugging (7.39) into (7.38), we conclude that

$$\begin{aligned}
&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\
&\qquad \le [2\, C_1^2\{s_{\mathrm{G}}\ln(p^2)/n\} + 2\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{\ln(12p)/n\}^{1/2} \\
&\qquad\qquad \times (4s_{\mathrm{G}}/\sqrt{k} + \sqrt{k})] \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\boldsymbol{\Delta}\|_{\ell_2}^2.
\end{aligned}$$

Now, to balance the terms $4\,s_{\mathrm{G}}/\sqrt{k}$ and $\sqrt{k}$ in the last line above, we choose $k = 4\,s_{\mathrm{G}}$. Then (B.1) translates into (3.8) with s_ℓ replaced by s_{G}, which holds by assumption, and therefore we have

$$\begin{aligned}
&|\langle \boldsymbol{\Delta}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\boldsymbol{\Delta}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\rangle| \\
&\qquad \le \left[2\, C_1^2\{s_{\mathrm{G}}\ln(p^2)/n\} + 8\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{s_{\mathrm{G}}\ln(12p)/n\}^{1/2}\right] \\
&\qquad\qquad \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \times \|\boldsymbol{\Delta}\|_{\ell_2}^2.
\end{aligned} \tag{7.40}$$

Then, from (7.34), (7.35), (7.17) and (7.40), we have

$$\begin{aligned}
&\langle \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^* + \boldsymbol{\Delta}; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}) - \nabla\mathcal{L}(\boldsymbol{\Delta}, \mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}), \boldsymbol{\Delta}\rangle \\
&\qquad \ge (2\kappa_{\mathrm{G}} - \|\boldsymbol{\Sigma}_{\mathbf{XX}}\|_{\mathrm{op}} \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}} - \boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}} \\
&\qquad\qquad - \left[2C_1^2\{s_{\mathrm{G}}\ln(p^2)/n\} + 8\,\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})\{s_{\mathrm{G}}\ln(12p)/n\}^{1/2}\right] \\
&\qquad\qquad\qquad \times \|\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}}\|_{\mathrm{op}})\|\boldsymbol{\Delta}\|_{\ell_2}^2.
\end{aligned} \tag{7.41}$$

Finally, invoking (4.6) in (7.41) yields the conclusion of the proposition. □

We wish to apply Lemma A.2. To this end, we check the three itemized conditions at the beginning of the proof of Proposition 3.1, except that now we check that appropriate RSC-D condition, instead of RSC condition, holds. We focus on the event $E_{\infty,1,n} \cap E_{\infty,2,n} \cap E_{\infty,3,n} \cap E_{\mathrm{op},4s_{\mathrm{G}},n} \cap F_{1,n} \cap F_{2,n} \cap F_{3,n}$ whose probability is at least $1 - (1/p + 1/q)^2 - 9/p$. Here the event $F_{1,n}$ is introduced in Proposition 7.5, and the events $F_{2,n}$ and $F_{3,n}$ are introduced below Proposition 7.6.

First, we set $\mathcal{M}$ and $\overline{\mathcal{M}}$ as in (7.28) again. Then again $\mathbf{B}^* \in \mathcal{M}$, the penalty $\mathcal{R}_{\mathrm{G}}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, and the subspace compatibility constant is $\Psi(\overline{\mathcal{M}}) = \sqrt{s_{\mathrm{G}}}$. Next, by Proposition 7.7 and assumption on $\kappa'_{\mathrm{D,G}}$, over the cone set $\mathbb{C}_{\mathrm{G}}(1)$, the loss $\mathcal{L}(\cdot; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})$ satisfies RSC-D condition (7.8) with tolerance function equal to zero and curvature $\kappa''_{\mathrm{D,G}} = \kappa_{\mathrm{D,G}} > 0$. Finally, essentially by (7.33), we can verify that

$$\mathcal{R}_{\mathrm{G}}^*\{\nabla\mathcal{L}(\mathbf{B}^*; \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}}, \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\varepsilon\varepsilon}})\} \le \lambda_{\mathrm{D,G}}.$$

The conclusions of the theorem then follow from Lemma A.2. □

8. Discussion

In this paper, we studied the estimation of the coefficient matrix in an elliptical copula multivariate response regression model. In this model, the joint distribution of the responses and covariates has an elliptical copula and arbitrary marginal distributions, and after applying some unknown monotonic marginal transformations, the responses and covariates are jointly elliptically distributed and are linked to one another in a linear regression model. As such, this model is much more flexible than the conventional multivariate response linear regression model. We provide penalized estimators of the coefficient matrix that in particular are computationally efficient and adaptive to the unknown marginal transformations, incorporate the precision matrix of the error vector and can take advantage of the potential row-sparsity of the coefficient matrix.

Extensions are possible. First, we could consider the issue of support recovery of $\mathbf{B}^*$, which we have not dealt with here. In particular, analogous to the improvement in the estimation of $\mathbf{B}^*$ that we have seen by employing a group penalty under a row-sparse model for $\mathbf{B}^*$, it would be interesting to investigate the extent to which group penalty will help with the recovery of row-support of $\mathbf{B}^*$ in our copula context. Second, our current algorithm terminates after an improved estimation of $\mathbf{B}^*$ is obtained. To further improve the estimation accuracy of $\mathbf{B}^*$ and $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon\varepsilon}}$, we could attempt to reiterate the procedures in Sections 4–5, but starting with the improved estimator of $\mathbf{B}^*$ instead of $\widetilde{\mathbf{B}}$ in (4.1). Convergence of the iteration scheme analogous to that derived in [22] is then needed.

Acknowledgments

The authors are grateful to Yi Yang for some useful discussions and suggestions. Funding for this work was provided by the Canada Research Chairs Program and grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

Appendix A: Results for the general Dantzig selector problem

We follow the convention set in Section 7.1. We let $\mathbb{Q}$ be a generic Euclidean space, and consider two subspaces $\mathcal{M} \subset \overline{\mathcal{M}}$ of $\mathbb{Q}$. Assume that the norm-based penalty function $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, let $\mathcal{R}^*$ denote the dual of $\mathcal{R}$, and let $\Psi(\mathcal{M})$ be the subspace compatibility constant with respect to the pair $(\mathcal{R}, \|\cdot\|_{\ell_2})$. We define the cone set

$$\mathbb{C} = \{\boldsymbol{\theta} \in \mathbb{Q} : \mathcal{R}(\boldsymbol{\theta}_{\overline{\mathcal{M}}^{\perp}}) \leq \mathcal{R}(\boldsymbol{\theta}_{\overline{\mathcal{M}}})\}\,.$$

Let $\mathcal{L} : \mathbb{Q} \to \mathbb{R}$ be a generic loss function and let $\boldsymbol{\beta}^*$ denote the true (but unknown) parameter value. To estimate $\boldsymbol{\beta}^*$, consider the solution $\widehat{\boldsymbol{\beta}}$ to the Dantzig selector program

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{Q}} \mathcal{R}(\boldsymbol{\beta}),$$

subject to $\mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta})\} \leq \lambda_{\mathrm{D}}$, where λ_{D} is a tuning parameter. Further let $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$.

The first lemma below is a slight extension of a well-known result on the geometry of the solution of a Dantzig selector program; see, e.g., Section III in [9]. The second lemma states bounds on $\boldsymbol{\delta}$.

Lemma A.1. *Assume that $\boldsymbol{\beta}^* \in \mathcal{M}$ (so $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*_{\mathcal{M}}$ and $\boldsymbol{\beta}^*_{\mathcal{M}^\perp} = \mathbf{0}$) and that $\boldsymbol{\beta}^*$ is a feasible solution, i.e., $\lambda_{\mathrm{D}} \geq \mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\}$. Then $\boldsymbol{\delta}$ always satisfies the cone condition $\boldsymbol{\delta} \in \mathbb{C}$.*

Proof. The derivation is standard. Here we follow the first part of the proof of Theorem 4.3 in [21]. By the triangle inequality and the assumption that $\boldsymbol{\beta}^* \in \mathcal{M}$, we have

$$\mathcal{R}(\widehat{\boldsymbol{\beta}}) = \mathcal{R}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) = \mathcal{R}(\boldsymbol{\beta}^*_{\mathcal{M}} + \boldsymbol{\delta}_{\overline{\mathcal{M}}} + \boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) \geq \mathcal{R}(\boldsymbol{\beta}^*_{\mathcal{M}} + \boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) - \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}}).$$

From the decomposability of $\mathcal{R}$ and the assumption $\boldsymbol{\beta}^* \in \mathcal{M}$, we also have

$$\mathcal{R}(\boldsymbol{\beta}^*_{\mathcal{M}} + \boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) = \mathcal{R}(\boldsymbol{\beta}^*_{\mathcal{M}}) + \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) = \mathcal{R}(\boldsymbol{\beta}^*) + \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}).$$

Therefore, $\mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) \leq \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}}) + \mathcal{R}(\widehat{\boldsymbol{\beta}}) - \mathcal{R}(\boldsymbol{\beta}^*)$. Furthermore, given that $\boldsymbol{\beta}^*$ is a feasible solution, we have $\mathcal{R}(\widehat{\boldsymbol{\beta}}) \leq \mathcal{R}(\boldsymbol{\beta}^*)$ and hence we can conclude. □

Lemma A.2. *Assume that $\boldsymbol{\beta}^* \in \mathcal{M}$ and that $\lambda_{\mathrm{D}} \geq \mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\}$. Further assume that $\mathcal{L}$ satisfies RSC-D condition (7.8) with curvature $\kappa_{\mathcal{F}} = \kappa_1 > 0$ and tolerance function $\tau_{\mathcal{F}}(\boldsymbol{\beta}^*) = \kappa_2\mathcal{R}^2(\boldsymbol{\delta})$ with $\kappa_2 \geq 0$ over the cone set $\mathbb{C}$, and that $\kappa_1 - 4\Psi^2(\mathcal{M})\kappa_2 > 0$. Then*

$$\|\boldsymbol{\delta}\|_{\ell_2} \leq 4\{\kappa_1 - 4\,\Psi^2(\mathcal{M})\kappa_2\}^{-1}\Psi(\mathcal{M})\lambda_{\mathrm{D}}, \tag{A.1}$$

and

$$\mathcal{R}(\boldsymbol{\delta}) \leq 8\{\kappa_1 - 4\,\Psi^2(\mathcal{M})\kappa_2\}^{-1}\Psi^2(\mathcal{M})\lambda_{\mathrm{D}}. \tag{A.2}$$

Proof. First observe that the assumption on λ_{D} makes $\boldsymbol{\beta}^*$ a feasible solution, so that $\boldsymbol{\delta} \in \mathbb{C}$ by Lemma A.1. Next note that by Hölder's inequality and the triangle inequality,

$$\begin{aligned}\langle \nabla\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle &\leq \mathcal{R}(\boldsymbol{\delta})\mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*)\} \\ &\leq \mathcal{R}(\boldsymbol{\delta})\left[\mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\delta})\} + \mathcal{R}^*\{\nabla\mathcal{L}(\boldsymbol{\beta}^*)\}\right].\end{aligned}$$

The assumption on λ_{D} and the fact that $\widehat{\boldsymbol{\beta}}$ is a feasible solution then imply that the right-hand side is bounded above by $2\lambda_{\mathrm{D}}\mathcal{R}(\boldsymbol{\delta})$. Furthermore, the triangle inequality, the fact that $\boldsymbol{\delta} \in \mathbb{C}$, and the definition of the subspace compatibility constant successively imply that

$$\begin{aligned}\mathcal{R}(\boldsymbol{\delta}) &\leq \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}^\perp}) + \mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}}) \leq 2\,\mathcal{R}(\boldsymbol{\delta}_{\overline{\mathcal{M}}}) \\ &\leq 2\,\Psi(\mathcal{M})\|\boldsymbol{\delta}_{\overline{\mathcal{M}}}\|_{\ell_2} \leq 2\,\Psi(\mathcal{M})\|\boldsymbol{\delta}\|_{\ell_2}.\end{aligned} \tag{A.3}$$

Therefore,

$$\langle \nabla\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle \leq 2\lambda_{\mathrm{D}}\mathcal{R}(\boldsymbol{\delta}) \leq 4\lambda_{\mathrm{D}}\Psi(\mathcal{M})\|\boldsymbol{\delta}\|_{\ell_2}. \tag{A.4}$$

Moreover, inequality (A.3) and the condition (7.8) on $\mathcal{L}$ imply

$$\begin{aligned}\langle \nabla\mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{\delta}) - \nabla\mathcal{L}(\boldsymbol{\beta}^*), \boldsymbol{\delta}\rangle &\geq \kappa_1\|\boldsymbol{\delta}\|^2_{\ell_2} - \kappa_2\mathcal{R}^2(\boldsymbol{\delta}) \\ &\geq \{\kappa_1 - 4\,\Psi^2(\mathcal{M})\kappa_2\}\|\boldsymbol{\delta}\|^2_{\ell_2}.\end{aligned} \tag{A.5}$$

Inequality (A.1) then follows from (A.4) and (A.5). Finally, (A.1) and (A.3) together yield (A.2). □

Appendix B: Other auxiliary lemmas

B.1. Deviation bound in operator norm for the plug-in estimator based on Kendall's tau

We propose a slight reformulation of Corollary 4.8 from [1].

Lemma B.1. *Let $k \in \mathbb{N}$ be an arbitrary integer and fix $\delta > 0$. Assume that*

$$n \geq \ln(2/\delta) + 2k \ln(12p). \tag{B.1}$$

Then there exists an event $E_{\mathrm{op},k,\delta,n}$ with probability at least $1 - \delta$ such that on the event $E_{\infty,1,n} \cap E_{\mathrm{op},k,\delta,n}$, we have, for all $\mathbf{u} \in \mathbb{R}^p$,

$$\begin{aligned} |\mathbf{u}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{u}| &\leq C_1^2 \ln(p^2)\, \|\mathbf{u}\|_{\ell_1}^2/(2n) \\ &\quad + \mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}}) \sqrt{\{\ln(2/\delta) + 2k\ln(12p)\}/n}\, (\|\mathbf{u}\|_{\ell_2}^2 + \|\mathbf{u}\|_{\ell_1}^2/k), \end{aligned} \tag{B.2}$$

where $\mathcal{C}'(\boldsymbol{\Sigma}_{\mathbf{XX}})$ is as defined in (3.10).

Proof. Following Section 4.3 of [1], define $\mathcal{S}_k = \{\mathbf{v} \in \mathbb{R}^p : \|\mathbf{v}\|_{\ell_2} \leq 1, \|\mathbf{v}\|_{\ell_0} \leq k\}$. For arbitrary $\mathbf{u} \in \mathbb{R}^p$, it follows easily from Lemma 4.9 and the proof of Lemma 4.7 in [1] that

$$\begin{aligned} |\mathbf{u}^\top(\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} - \boldsymbol{\Sigma}_{\mathbf{XX}})\mathbf{u}| &\leq (\pi^2/8)\|\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}}\|_{\ell_\infty}^2\, \|\mathbf{u}\|_{\ell_1}^2 \\ &\quad + 2\pi \sup_{\mathbf{v},\mathbf{v}' \in \mathcal{S}_k} |\mathbf{v}^\top(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{v}'|\, (\|\mathbf{u}\|_{\ell_2} + \|\mathbf{u}\|_{\ell_1}/\sqrt{k})^2. \end{aligned} \tag{B.3}$$

Next, Lemma 4.6 in [1] also states that on an event $E_{\mathrm{op},k,\delta,n}$ with probability at least $1 - \delta$,

$$\sup_{\mathbf{v},\mathbf{v}' \in \mathcal{S}_k} |\mathbf{v}^\top(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{v}'| \leq 32(1+\sqrt{5})\, \mathcal{C}(\boldsymbol{\Sigma}_{\mathbf{XX}}) \sqrt{\{\ln(2/\delta) + 2k\ln(12p)\}/n}.$$

Thus (B.2) follows from (B.3), the above bound on $E_{\mathrm{op},k,\delta,n}$, and the bound (7.1) on $E_{\infty,1,n}$. □

B.2. A matrix inequality

We have the following result.

Proposition B.2. *Consider an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$. Then*

$$\sum_{\ell=1}^{q} \|(\mathbf{A})_{\bullet\ell}\|_{\ell_1}^2 \leq \|\mathbf{A}\|_{\ell_1,\ell_2}^2. \tag{B.4}$$

Proof. Assume without loss of generality that $\mathbf{A}$ has entries $a_{k\ell} \geq 0$. Then

$$\sum_{\ell=1}^{q} \|(\mathbf{A})_{\bullet\ell}\|_{\ell_1}^2 = \sum_{\ell=1}^{q}\Big(\sum_{k=1}^{p} a_{k\ell}\Big)^2 = \sum_{\ell=1}^{q}\sum_{k,k'=1}^{p} a_{k\ell}a_{k'\ell} = \sum_{k,k'=1}^{p}\Big(\sum_{\ell=1}^{q} a_{k\ell}a_{k'\ell}\Big)$$

and

$$\|\mathbf{A}\|_{\ell_1,\ell_2}^2 = \Big\{ \sum_{k=1}^{p} \Big(\sum_{\ell=1}^{q} a_{k\ell}^2 \Big)^{1/2} \Big\}^2 = \sum_{k,k'=1}^{p} \Big\{ \Big(\sum_{\ell=1}^{q} a_{k\ell}^2 \Big)^{1/2} \Big(\sum_{\ell=1}^{q} a_{k'\ell}^2 \Big)^{1/2} \Big\}.$$

Thus it suffices to show that for each pair $(k, k') \in [p] \times [p]$, we have

$$\Big(\sum_{\ell=1}^{q} a_{k\ell} a_{k'\ell} \Big)^2 \le \Big(\sum_{\ell=1}^{q} a_{k\ell}^2 \Big) \Big(\sum_{\ell=1}^{q} a_{k'\ell}^2 \Big),$$

which is an easy consequence of the Cauchy–Schwarz inequality. □

B.3. Conditional moments of Gaussian distributions

Lemma B.3. *Let Y and Z be $\mathcal{N}(0,1)$ random variables that are in addition jointly normal with correlation ρ. For any integer $r \in \mathbb{N}$,*

$$\mathrm{E}(|Y|^r | Z = z) \le 2^{r-1} \rho^r |z|^r + 2^{r-1} (1-\rho^2)^{r/2} (r-1)!!$$

where $r!!$ denotes the double or semifactorial of an integer r, i.e., the product of all the integers from 1 up to r that have the same parity (odd or even) as r.

Proof. Given that $Y|Z = z$ has the same distribution as $\rho z + (1-\rho^2)^{1/2} Y$, we can write

$$\begin{aligned} \mathrm{E}(|Y|^r | Z = z) &= \mathrm{E}\{|\rho z + (1-\rho^2)^{1/2} Y|^r\} \\ &\le 2^{r-1} \rho^r |z|^r + 2^{r-1} (1-\rho^2)^{r/2} \mathrm{E}(|Y|^r), \end{aligned}$$

from which the conclusion follows from the formula for the rth absolute moment of Y. □

B.4. Bernstein's inequality

The following result is Theorem 2.10 in [2].

Lemma B.4. *Let $X_1, \ldots, X_n$ be centered independent random variables such that for every integer $k \ge 2$ and all $i \in [n]$, $\mathrm{E}(|X_i|^k) \le k! \sigma_i^2 c^{k-2}/2$, and set*

$$\sigma^2 = \sum_{i \in [n]} \sigma_i^2, \quad S_n = \sum_{i \in [n]} X_i.$$

Then, for all $t \in [0, \infty)$, $\Pr(S_n \ge \sqrt{2\sigma^2 t} + ct) \le e^{-t}$.

Appendix C: Bounding the $\|\cdot\|_{\ell_\infty,\ell_2}$ norm of some random matrices

Let $\mathcal{S}^{q-1} = \{\mathbf{v} \in \mathbb{R}^q : \|\mathbf{v}\|_{\ell_2} = 1\}$ be the unit Euclidean sphere in $\mathbb{R}^q$ and let $\mathcal{N}$ be a (1/2)-net of $\mathcal{S}^{q-1}$ equipped with the Euclidean norm; see Definition 5.1 in [40]. From, e.g., Lemma 5.2 in [40], we can and will take $\mathcal{N}$ to satisfy $\mathrm{card}(\mathcal{N}) \leq \{1+1/(1/2)\}^q = 5^q$. Let $\boldsymbol{\iota}_k \in \mathbb{R}^p$ have a 1 at the kth coordinate, and 0 elsewhere. Then, for any matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$,

$$\|\mathbf{M}\|_{\ell_\infty,\ell_2} = \sup_{k\in[p]} \|\boldsymbol{\iota}_k^\top \mathbf{M}\|_{\ell_2} = \sup_{k\in[p],\mathbf{v}\in\mathcal{S}^{q-1}} \boldsymbol{\iota}_k^\top \mathbf{M}\mathbf{v}. \tag{C.1}$$

The following result shows that the supremum over $\mathbf{v} \in \mathcal{S}^{q-1}$ in the above can be replaced by a supremum over the net $\mathcal{N}$ at the cost of a constant factor.

Proposition C.1. *For any* $\mathbf{M} \in \mathbb{R}^{p\times q}$*, we have*

$$\|\mathbf{M}\|_{\ell_\infty,\ell_2} \leq 2 \sup_{k\in[p],\mathbf{v}\in\mathcal{N}} \boldsymbol{\iota}_k^\top \mathbf{M}\mathbf{v}.$$

Proof. In view of (C.1), we can find $k^* \in [p]$ and $\mathbf{v}^* \in \mathcal{S}^{q-1}$ such that $\|\mathbf{M}\|_{\ell_\infty,\ell_2} = \|\boldsymbol{\iota}_{k^*}^\top \mathbf{M}\|_{\ell_2} = \boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}^*$. Choose $\mathbf{v}' \in \mathcal{N}$ so that $\|\mathbf{v}^* - \mathbf{v}'\|_{\ell_2} \leq 1/2$. Then

$$|\boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}^* - \boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}'| \leq \|\boldsymbol{\iota}_{k^*}^\top \mathbf{M}\|_{\ell_2} \times \|\mathbf{v}^* - \mathbf{v}'\|_{\ell_2} \leq \|\mathbf{M}\|_{\ell_\infty,\ell_2}/2,$$

and hence $\boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}' \geq \boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}^* - \|\mathbf{M}\|_{\ell_\infty,\ell_2}/2 = \|\mathbf{M}\|_{\ell_\infty,\ell_2}/2$, which implies that $\|\mathbf{M}\|_{\ell_\infty,\ell_2} \leq 2\boldsymbol{\iota}_{k^*}^\top \mathbf{M}\mathbf{v}'$. This is enough to conclude. □

C.1. Proof of Proposition 7.5

First note that because the sine function is Lipschitz, we have

$$\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{XY}} - \boldsymbol{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2} \leq \pi \, \|\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2}/2.$$

So it suffices to bound the right-hand side from above. The proof proceeds in three steps.

C.1.1. Reduction to a net

By Proposition C.1, we have

$$\|\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2} \leq 2 \sup_{k\in[p],\mathbf{v}\in\mathcal{N}} \boldsymbol{\iota}_k^\top (\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}})\mathbf{v}.$$

For fixed $k \in [p]$, $\mathbf{v} \in \mathcal{N}$ and $\delta > 0$, consider the event

$$A_{k,\mathbf{v},\delta} = \{\boldsymbol{\iota}_k^\top (\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}})\mathbf{v} < \delta\}.$$

Using the Chernoff bound technique, we then have, for all $t > 0$,

$$\Pr\{(A_{k,\mathbf{v},\delta})^{\complement}\} \leq e^{-t\delta} \mathrm{E}\big[\exp\{t\, \boldsymbol{\iota}_k^\top (\widehat{\mathbf{T}}_{\mathbf{XY}} - \mathbf{T}_{\mathbf{XY}})\mathbf{v}\}\big]. \tag{C.2}$$

C.1.2. Decoupling the U-statistic

For arbitrary $i, j \in [n]$, let

$$h(\mathbf{X}_i, \mathbf{X}_j, \mathbf{Y}_i, \mathbf{Y}_j) = \operatorname{sgn}(\mathbf{X}_i - \mathbf{X}_j) \operatorname{sgn}(\mathbf{Y}_i - \mathbf{Y}_j)^\top,$$

and for any permutation $i_1, \ldots, i_n$ of the integers $1, \ldots, n$, define

$$\begin{aligned} W(&\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n}, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_n}) \\ &= 2\{h(\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \mathbf{Y}_{i_1}, \mathbf{Y}_{i_2}) + h(\mathbf{X}_{i_3}, \mathbf{X}_{i_4}, \mathbf{Y}_{i_3}, \mathbf{Y}_{i_4}) \\ &\qquad + \cdots + h(\mathbf{X}_{i_{n-1}}, \mathbf{X}_{i_n}, \mathbf{Y}_{i_{n-1}}, \mathbf{Y}_{i_n})\}/n. \end{aligned}$$

Summing over all possible permutations, we can then write

$$\widehat{\mathbf{T}}_{\mathbf{XY}} = \sum_{n,n} W(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n}, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_n})/n!$$

in terms of the sample $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$. It then follows from (C.2) and the convexity of the exponential function that

$$\begin{aligned} \Pr\{&(A_{k,\mathbf{v},\delta})^{\complement}\} \\ &\le e^{-t\delta} \sum_{n,n} \mathrm{E}\big[\exp[t\,\boldsymbol{\iota}_k^\top\{W(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n}, \mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_n}) - \mathbf{T}_{\mathbf{XY}}\}\mathbf{v}]\big]/n! \\ &= e^{-t\delta}\mathrm{E}\big[\exp[t\,\boldsymbol{\iota}_k^\top\{W(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \mathbf{T}_{\mathbf{XY}}\}\mathbf{v}]\big]. \end{aligned}$$

C.1.3. Conversion into an average of sub-Gaussian random variables

For each $i \in [n/2]$, let

$$\mathbf{S}_{\mathbf{X},i} = \operatorname{sgn}(\mathbf{X}_{2i-1} - \mathbf{X}_{2i}), \quad \mathbf{S}_{\mathbf{Y},i} = \operatorname{sgn}(\mathbf{Y}_{2i-1} - \mathbf{Y}_{2i}), \tag{C.3}$$

so that

$$\begin{aligned} \boldsymbol{\iota}_k^\top&\{W(\mathbf{X}_1, \ldots, \mathbf{X}_n, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) - \mathbf{T}_{\mathbf{XY}}\}\mathbf{v} \\ &= 2\boldsymbol{\iota}_k^\top \sum_{i=1}^{n/2} \{\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top - \mathrm{E}(\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top)\}\mathbf{v}/n, \end{aligned}$$

and hence

$$\Pr\{(A_{k,\mathbf{v},\delta})^{\complement}\} \le e^{-t\delta}\mathrm{E}\Big[\exp\Big[2t\boldsymbol{\iota}_k^\top \sum_{i=1}^{n/2} \{\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top - \mathrm{E}(\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top)\}\mathbf{v}/n\Big]\Big]. \tag{C.4}$$

For each $i \in [n/2]$, further let $U_{k,i} = \boldsymbol{\iota}_k^\top \mathbf{S}_{\mathbf{X},i}$ and $W_{\mathbf{v},i} = \mathbf{v}^\top \mathbf{S}_{\mathbf{Y},i}$. Then

$$\boldsymbol{\iota}_k^\top \sum_{i=1}^{n/2} \{\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top - \mathrm{E}(\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{Y},i}^\top)\}\mathbf{v} = \sum_{i=1}^{n/2} \{U_{k,i}W_{\mathbf{v},i} - \mathrm{E}(U_{k,i}W_{\mathbf{v},i})\}.$$

Given that the variables $\mathbf{S}_{\mathbf{X},1}, \ldots, \mathbf{S}_{\mathbf{X},n/2}$ are iid, as are $\mathbf{S}_{\mathbf{Y},1}, \ldots, \mathbf{S}_{\mathbf{Y},n/2}$, we have

$$\mathrm{E}\Big[\exp\Big[2t\sum_{i=1}^{n/2}\{U_{k,i}W_{\mathbf{v},i} - \mathrm{E}(U_{k,i}W_{\mathbf{v},i})\}/n\Big]\Big] \\ = \big[\mathrm{E}\exp[2t\{U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\}/n]\big]^{n/2},$$

and hence (C.4) becomes

$$\Pr\{(A_{k,\mathbf{v},\delta})^{\complement}\} \le e^{-t\delta}\big[\mathrm{E}\exp[2t\{U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\}/n]\big]^{n/2}. \qquad \text{(C.5)}$$

The problem is thus reduced to finding an upper bound on the moment generating function of $U_{k,1}W_{\mathbf{v},1}$. To this end, define the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ of a random variable R by

$$\|R\|_{\psi_2} = \sup_{m\in\mathbb{N}} m^{-1/2}\left(\mathrm{E}|R|^m\right)^{1/m},$$

as in Definition 5.7 of [40]. By Remark 5.18 in [40], and considering that $|U_{k,1}| = 1$, we have

$$\|U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\|_{\psi_2} \le 2\,\|U_{k,1}W_{\mathbf{v},1}\|_{\psi_2} = 2\,\|W_{\mathbf{v},1}\|_{\psi_2}. \qquad \text{(C.6)}$$

Furthermore, by Lemma 5.5 in [40], notably the equivalence of the constants K_2 and K_4 in that lemma up to an absolute constant factor, there exists an absolute constant C' such that

$$\begin{aligned}\mathrm{E}\exp\big[2t\{U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\}/n\big] &\\ \le \exp\{C't^2\|U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\|^2_{\psi_2}/n^2\} &\\ \le \exp(C'4t^2\|W_{\mathbf{v},1}\|^2_{\psi_2}/n^2), &\qquad \text{(C.7)}\end{aligned}$$

where the second inequality follows by (C.6).

Next, Lemmas 4.4–4.5 in [1] imply that for all $i \in [n/2]$, $\mathbf{S}_{\mathbf{Y},i}$ is $\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})$-sub-Gaussian, i.e., for any fixed $\mathbf{w} \in \mathbb{R}^q$, we have

$$\mathrm{E}\exp(\mathbf{w}^\top\mathbf{S}_{\mathbf{Y},i}) \le \exp\{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})\,\|\mathbf{w}\|^2_{\ell^2}/2\}.$$

Thus, for all $\theta \in \mathbb{R}$,

$$\begin{aligned}\mathrm{E}\{\exp(\theta W_{\mathbf{v},1})\} &= \mathrm{E}\exp\{(\theta\mathbf{v})^\top\mathbf{S}_{\mathbf{Y},1}\}\\ &\le \exp\{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})\,\|\theta\mathbf{v}\|^2_{\ell^2}/2\} = \exp\{\theta^2\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})/2\}.\end{aligned}$$

By the above and again by Lemma 5.5 in [40], in particular the equivalence of the constants K_2 and K_4 in that lemma up to an absolute constant factor, the centered random variable $W_{\mathbf{v},1}$ satisfies

$$\|W_{\mathbf{v},1}\|_{\psi_2} \le C''\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})} \qquad \text{(C.8)}$$

for some absolute constant C''. Combining (C.7) and (C.8), we conclude that

$$\mathrm{E}\exp[2t\{U_{k,1}W_{\mathbf{v},1} - \mathrm{E}(U_{k,1}W_{\mathbf{v},1})\}/n] \leq \exp\{t^2 C''' \,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})/n^2\}, \tag{C.9}$$

with $C''' = (4C') \times (C'')^2$. Combining (C.9) and (C.5), we get

$$\Pr\{(A_{k,\mathbf{v},\delta})^{\complement}\} \leq \exp\{-t\delta + t^2 C''' \,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})/(2n)\}$$

for all $t \in (0,\infty)$, and hence also, by minimizing over all such values of t,

$$\Pr\{(A_{k,\mathbf{v},\delta})^{\complement}\} \leq \exp\big[-n\delta^2/\{2C'''\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})\}\big].$$

Thus if

$$\delta = \delta^* = \sqrt{2C''' \,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})}\,\sqrt{\{\ln(p^2) + \ln(5)q\}/n},$$

we then hav e $\Pr\{(A_{k,\mathbf{v},\delta^*})^{\complement}\} \leq 1/(p^2\,5^q)$. Finally, consider the event $F_{1,n} = \cap_{k\in[p],\mathbf{v}\in\mathcal{N}} A_{k,\mathbf{v},\delta^*}$. From the above upper bound on $\Pr\{(A_{k,\mathbf{v},\delta^*})^{\complement}\}$, valid for all $k \in [p]$ and $\mathbf{v} \in \mathcal{N}$, together with the union bound, we get $\Pr(F_{1,n}) \geq 1 - 1/p$. Summing up, we have

$$\|\widehat{\mathbf{\Sigma}}_{\mathbf{XY}} - \mathbf{\Sigma}_{\mathbf{XY}}\|_{\ell_\infty,\ell_2} \leq \pi\delta^* = \pi\,\sqrt{2C''' \,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{YY}})}\,\sqrt{\{\ln(p^2) + \ln(5)q\}/n},$$

on the event $F_{1,n}$, which is the desired conclusion. □

C.2. Proof of Proposition 7.6

It mimics the proof of Proposition 7.5 but extra steps are needed to tackle the right-multiplication of $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}}$ by $\mathbf{D}$. First, a Taylor expansion yields

$$\begin{aligned}\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2} &= \|(\pi/2)\cos(\pi\mathbf{T}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{D} \\ &\quad - (\pi^2/8)\sin(\pi\overline{\mathbf{T}}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2},\end{aligned}$$

which, by the triangle inequality, is bounded above by

$$\begin{aligned}&(\pi/2)\|\cos(\pi\mathbf{T}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2} \\ &\quad + (\pi^2/8)\|\sin(\pi\overline{\mathbf{T}}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2},\end{aligned} \tag{C.10}$$

where $\overline{\mathbf{T}}_{\mathbf{XX}}$ is a symmetric random matrix such that each entry $(\overline{\mathbf{T}}_{\mathbf{XX}})_{k\ell}$ is a random number strictly between $(\widehat{\mathbf{T}}_{\mathbf{XX}})_{k\ell}$ and $(\mathbf{T}_{\mathbf{XX}})_{k\ell}$. The second summand in (C.10) is bounded above by

$$\begin{aligned}&(\pi^2\sqrt{q}/8)\|\sin(\pi\overline{\mathbf{T}}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty} \\ &\quad \leq (\pi^2\sqrt{q}/8)\,\|\sin(\pi\overline{\mathbf{T}}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\circ(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\|_{\ell_\infty} \times \|\mathbf{D}\|_1 \\ &\quad \leq (\pi^2\sqrt{q}/8)\,\|\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}}\|^2_{\ell_\infty} \times \|\mathbf{D}\|_1.\end{aligned} \tag{C.11}$$

It remains to find an upper bound on the first summand in (C.10).

C.2.1. Reduction to a net

By Proposition C.1, the first summand in (C.10) is no larger than

$$\pi \sup_{k\in[p],\mathbf{v}\in\mathcal{N}} \boldsymbol{\iota}_k^\top \cos(\pi\mathbf{T}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathbf{D}\mathbf{v}.$$

For fixed $k\in[p]$, $\mathbf{v}\in\mathcal{N}$ and $\delta>0$, consider the event

$$B_{k,\mathbf{v},\delta}=\{\boldsymbol{\iota}_k^\top \cos(\pi\mathbf{T}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathbf{D}\mathbf{v}<\delta\}.$$

Using the Chernoff bound technique, we have, for all $t>0$,

$$\Pr\{(B_{k,\mathbf{v},\delta})^{\complement}\}\le e^{-t\delta}\mathrm{E}\big[\exp\{t\boldsymbol{\iota}_k^\top \cos(\pi\mathbf{T}_{\mathbf{XX}}/2)\circ(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathbf{D}\mathbf{v}\}\big]. \quad \text{(C.12)}$$

C.2.2. Treating the cosine function transformation

By Lemma E.1 in [1], there exist vectors $\mathbf{a}_1,\mathbf{a}_2,\ldots$ and $\mathbf{b}_1,\mathbf{b}_2,\ldots$, all belonging to $\mathbb{R}^p$, with $\|\mathbf{a}_r\|_{\ell_\infty},\|\mathbf{b}_r\|_{\ell_\infty}\le 1$ for every integer $r\in\mathbb{N}$, and a sequence $t_1,t_2.\ldots$ of non-negative numbers adding up to 4, such that

$$\cos(\pi\mathbf{T}_{\mathbf{XX}}/2)=\sum_{r\in\mathbb{N}} t_r\mathbf{a}_r\mathbf{b}_r^\top.$$

For any vector $\mathbf{u}$, let $\mathrm{diag}(\mathbf{u})$ be the diagonal matrix with the elements of $\mathbf{u}$ arranged on the diagonal. Plugging the above expression into the expectation term on the right-hand side of (C.12), and calling on the fact that the exponential function is convex, we find

$$\begin{aligned}\Pr\{(B_{k,\mathbf{v},\delta})^{\complement}\}&\le e^{-t\delta}\mathrm{E}\Big[\exp\Big\{\sum_{r\in\mathbb{N}} t_r\, t\,\boldsymbol{\iota}_k^\top\mathrm{diag}(\mathbf{a}_r)(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathrm{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\Big\}\Big]\\ &\le e^{-t\delta}\sum_{r\in\mathbb{N}} t_r\,\mathrm{E}\big[\exp\{t\,\boldsymbol{\iota}_k^\top\mathrm{diag}(\mathbf{a}_r)(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathrm{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\}\big],\end{aligned}$$

and hence

$$\begin{aligned}&\Pr\{(B_{k,\mathbf{v},\delta})^{\complement}\}\\ &\qquad\le 4e^{-t\delta}\max_{r\in\mathbb{N}}\mathrm{E}\big[\exp\{t\,\boldsymbol{\iota}_k^\top\mathrm{diag}(\mathbf{a}_r)(\widehat{\mathbf{T}}_{\mathbf{XX}}-\mathbf{T}_{\mathbf{XX}})\mathrm{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\}\big]. \quad \text{(C.13)}\end{aligned}$$

C.2.3. Decoupling the U-statistic

For arbitrary $i,j\in[n]$, let

$$h(\mathbf{X}_i,\mathbf{X}_j)=\mathrm{sgn}(\mathbf{X}_i-\mathbf{X}_j)\,\mathrm{sgn}(\mathbf{X}_i-\mathbf{X}_j)^\top,$$

and for any permutation $i_1,\ldots,i_n$ of the integers $1,\ldots,n$, write

$$V(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_n})=2\{h(\mathbf{X}_{i_1},\mathbf{X}_{i_2})+h(\mathbf{X}_{i_3},\mathbf{X}_{i_4})+\cdots+h(\mathbf{X}_{i_{n-1}},\mathbf{X}_{i_n})\}/n.$$

Summing over all possible permutations, we can again write

$$\widehat{\mathbf{T}}_{\mathbf{XX}} = \sum_{n,n} V(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n})/n!.$$

Thus, for any integer $r \in \mathbb{N}$,

$$\begin{aligned} &\mathrm{E} \exp \big\{(t\, \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\big\} \\ &\quad = \mathrm{E} \exp \Big[t\, \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)\Big[\sum_{n,n}\{V(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n}) - \mathbf{T}_{\mathbf{XX}}\}/n!\Big]\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\Big], \end{aligned}$$

and from the convexity of the exponential function, this expression is bounded above by

$$\begin{aligned} &\sum_{n,n} \mathrm{E}\big[\exp[t\, \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)\{V(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_n}) - \mathbf{T}_{\mathbf{XX}}\}\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}]\big]/n! \\ &\qquad\qquad = \mathrm{E}\big[\exp[t\, \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)\{V(\mathbf{X}_1, \ldots, \mathbf{X}_n) - \mathbf{T}_{\mathbf{XX}}\}\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}]\big]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathrm{E} \exp\{(t\, \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)(\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}})\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\} \\ &\qquad \le \mathrm{E}\Big[\exp\Big[2t\boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r) \\ &\qquad\qquad \times \sum_{i\in[n/2]} \{\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{X},i}^\top - \mathrm{E}(\mathbf{S}_{\mathbf{X},i}\mathbf{S}_{\mathbf{X},i}^\top)\}\operatorname{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}/n\Big]\Big], \end{aligned} \tag{C.14}$$

where $\mathbf{S}_{\mathbf{X},1}, \ldots, \mathbf{S}_{\mathbf{X},n/2}$ are as defined in (C.3) and iid.

C.2.4. Conversion into an average of sub-Gaussian random variables

For each $i \in [n/2]$, define $U_{k,r,i} = \boldsymbol{\iota}_k^\top \operatorname{diag}(\mathbf{a}_r)\mathbf{S}_{\mathbf{X},i}$ and $V_{\mathbf{v},r,i} = \mathbf{v}^\top\mathbf{D}^\top \operatorname{diag}(\mathbf{b}_r)\mathbf{S}_{\mathbf{X},i}$. Given that the pairs $(U_{k,r,i}, V_{\mathbf{v},r,i})$ are iid, we can rewrite the right-hand side of (C.14) as

$$\begin{aligned} &\mathrm{E}\Big[\exp\Big[2t\sum_{i\in[n/2]}\{U_{k,r,i}V_{\mathbf{v},r,i} - \mathrm{E}(U_{k,r,i}V_{\mathbf{v},r,i})\}/n\Big]\Big] \\ &\qquad\qquad = \big[\mathrm{E}\exp[2t\{U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\}/n]\big]^{n/2}. \end{aligned}$$

It then follows from (C.13) that

$$\begin{aligned} &\Pr\{(B_{k,\mathbf{v},\delta})^{\complement}\} \\ &\qquad \le 4\, e^{-t\delta} \max_{r\in\mathbb{N}} \mathrm{E}\big[\exp[2t\{U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\}/n]\big]^{n/2}. \end{aligned} \tag{C.15}$$

The problem is thus reduced to finding an upper bound on the moment generating function of $U_{k,r,1}V_{\mathbf{v},r,1}$. To this end, we use its sub-Gaussian property.

Given that $\|\mathbf{a}_r\|_{\ell_\infty} \le 1$, we have $|U_{k,r,1}| \le 1$ and hence $\|U_{k,r,1}V_{\mathbf{v},r,1}\|_{\psi_2} \le \|V_{\mathbf{v},r,1}\|_{\psi_2}$. By Remark 5.18 in [40], we also have

$$\|U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\|_{\psi_2} \le 2\,\|U_{k,r,1}V_{\mathbf{v},r,1}\|_{\psi_2} \le 2\,\|V_{\mathbf{v},r,1}\|_{\psi_2}. \quad \text{(C.16)}$$

By Lemma 5.5 in [40], and specifically the equivalence of the constants K_2 and K_4 in that lemma up to an absolute constant factor, there exists an absolute constant C' as introduced in the proof of Proposition 7.5, such that

$$\begin{aligned} &\mathrm{E}\big[\exp[2t\{U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\}/n]\big] \\ &\qquad\qquad \le \exp\{C't^2\|U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\|_{\psi_2}^2/n^2\} \\ &\qquad\qquad\qquad\qquad \le \exp(4C't^2\|V_{\mathbf{v},r,1}\|_{\psi_2}^2/n^2), \quad \text{(C.17)} \end{aligned}$$

where the second inequality follows by (C.16).

Next, Lemmas 4.4–4.5 in [1] imply that for all $i \in [n/2]$, $\mathbf{S}_{\mathbf{X},i}$ is $\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})$-sub-Gaussian, i.e., for any fixed $\mathbf{u} \in \mathbb{R}^p$, we have

$$\mathrm{E}\exp(\mathbf{u}^\top \mathbf{S}_{\mathbf{X},i}) \le \exp\{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})\,\|\mathbf{u}\|_{\ell^2}^2/2\}.$$

Thus, for all $\theta \in \mathbb{R}$,

$$\begin{aligned} \mathrm{E}\{\exp(\theta V_{\mathbf{v},r,1})\} &= \mathrm{E}\big[\exp[\{\theta \mathrm{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\}^\top \mathbf{S}_{\mathbf{X},1}]\big] \\ &\le \exp\{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})\,\|\theta \mathrm{diag}(\mathbf{b}_r)\mathbf{D}\mathbf{v}\|_{\ell^2}^2/2\} \\ &\le \exp\{\theta^2\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})\,\|\mathbf{D}\|_{\mathrm{op}}^2/2\}. \end{aligned}$$

From the above and again by Lemma 5.5 in [40], in particular the equivalence of the constants K_2 and K_4 in that lemma up to an absolute constant factor, the centered random variable $V_{\mathbf{v},r,1}$ satisfies

$$\|V_{\mathbf{v},r,1}\|_{\psi_2} \le C''\,\|\mathbf{D}\|_{\mathrm{op}}\sqrt{\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}, \quad \text{(C.18)}$$

for the same constant C'' as in (C.8). Combining (C.17) and (C.18), we conclude that

$$\begin{aligned} &\mathrm{E}\big[\exp[2t\{U_{k,r,1}V_{\mathbf{v},r,1} - \mathrm{E}(U_{k,r,1}V_{\mathbf{v},r,1})\}/n]\big] \\ &\qquad\qquad\qquad \le \exp\{C'''\,t^2\|\mathbf{D}\|_{\mathrm{op}}^2\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})/n^2\}, \quad \text{(C.19)} \end{aligned}$$

for the same constant C''' as in (C.9). In view of (C.19), (C.15) thus becomes

$$\Pr\{(B_{k,\mathbf{v},\delta})^\complement\} \le 4\exp\{-t\delta + t^2C'''\,\|\mathbf{D}\|_{\mathrm{op}}^2\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})/(2n)\}$$

for all $t \in (0,\infty)$, and hence also, by minimizing over all such values of t,

$$\Pr\{(B_{k,\mathbf{v},\delta})^\complement\} \le 4\exp[-n\delta^2/\{2C'''\|\mathbf{D}\|_{\mathrm{op}}^2\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})\}].$$

Thus $\Pr\{(B_{k,\mathbf{v},\delta^*})^\complement\} \le 4/(p^2 5^q)$ as soon as

$$\delta = \delta^* = \sqrt{2C'''\,\|\mathbf{D}\|_{\mathrm{op}}^2\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\,\sqrt{\{\ln(p^2) + \ln(5)q\}/n}.$$

Finally, consider the event $F_{\mathbf{D},n} = \cap_{k\in[p],\mathbf{v}\in\mathcal{N}} B_{k,\mathbf{v},\delta^*}$. From the above upper bound on $\Pr\{(B_{k,\mathbf{v},\delta^*})^{\complement}\}$, valid for all $k \in [p]$ and $\mathbf{v} \in \mathcal{N}$, together with the union bound, we get $\Pr(F_{\mathbf{D},n}) \geq 1 - 4/p$. Summing up (in particular recall (C.11)), we have, on the event $E_{\infty,1,n} \cap F_{\mathbf{D},n}$

$$\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2} \leq \pi\delta^* + (\pi^2\sqrt{q}/8)\|\widehat{\mathbf{T}}_{\mathbf{XX}} - \mathbf{T}_{\mathbf{XX}}\|_{\ell_\infty}^2 \|\mathbf{D}\|_1,$$

and hence, by invoking (7.1), we have, on the same event,

$$\begin{aligned}&\|(\widehat{\mathbf{\Sigma}}_{\mathbf{XX}} - \mathbf{\Sigma}_{\mathbf{XX}})\mathbf{D}\|_{\ell_\infty,\ell_2} \\ &\leq \pi\|\mathbf{D}\|_{\mathrm{op}}\sqrt{2C'''\,\mathcal{C}(\mathbf{\Sigma}_{\mathbf{XX}})}\sqrt{\{\ln(p^2)+\ln(5)q\}/n} + C_1^2\|\mathbf{D}\|_1\sqrt{q}\,\ln(p^2)/(2n),\end{aligned}$$

which is the desired conclusion. □

Appendix D: Computational aspects

We can write (5.1) in a form to which a solver for a standard Lasso problem can be readily applied. Section 2 of [6], specifically at the top of p. 6 around Eq. (2.2), describes a procedure similar in spirit but that applies to a univariate response problem. To see how, first recall that $\mathbf{S}$, $\mathbf{S}_\times$ and $\mathbf{\Omega}$ are supposed to be proxies for $\mathbf{\Sigma}_{\mathbf{XX}}$, $\mathbf{\Sigma}_{\mathbf{XY}}$ and $\mathbf{\Omega}_{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}}$, respectively. We further assume that $\mathbf{S}$ and $\mathbf{\Omega}$ are positive definite. Let $\mathbf{A} \in \mathbb{R}^{(pq)\times(pq)}$ be a Cholesky factor of $\mathbf{\Omega}\otimes\mathbf{S}$, so that $\mathbf{A}^\top\mathbf{A} = \mathbf{\Omega}\otimes\mathbf{S}$. Next, let $\mathbf{y} \in \mathbb{R}^{(pq)}$ be such that $\mathbf{A}^\top\mathbf{y} = \mathrm{vec}(\mathbf{S}_\times\mathbf{\Omega})$. We can then rewrite the right-hand side of (5.1) as

$$\begin{aligned}&\mathrm{vec}(\mathbf{B})^\top\mathbf{\Omega}\otimes\mathbf{S}\,\mathrm{vec}(\mathbf{B})/2 - \mathrm{tr}(\mathbf{\Omega}\mathbf{S}_\times^\top\mathbf{B}) \\ &\qquad = \mathrm{vec}(\mathbf{B})^\top\mathbf{A}^\top\mathbf{A}\,\mathrm{vec}(\mathbf{B})/2 - \mathrm{vec}(\mathbf{S}_\times\mathbf{\Omega})^\top\,\mathrm{vec}(\mathbf{B}) \\ &\qquad = \mathrm{vec}(\mathbf{B})^\top\mathbf{A}^\top\mathbf{A}\,\mathrm{vec}(\mathbf{B})/2 - \mathbf{y}^\top\mathbf{A}\,\mathrm{vec}(\mathbf{B}) \\ &\qquad = \|\mathbf{y} - \mathbf{A}\,\mathrm{vec}(\mathbf{B})\|_{\ell_2}^2/2 - \mathbf{y}^\top\mathbf{y}/2. \qquad \text{(D.1)}\end{aligned}$$

We have thus converted (5.1) to an equivalent univariate response least squares form (D.1) with $\mathbf{y}$ as the response, $\mathbf{A}$ as the explanatory variables, and the vectorized $\mathbf{B}$ as the regression coefficient.

Appendix E: Identifiability issues and prediction

Suppose we are given Model (1.1) but not all components of $\mathbf{X}$ and $\mathbf{Y}$ have mean 0 and variance 1. The model can then be transformed into a form that does. Model (1.1) implies that

$$\begin{aligned}&\{\mathrm{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}(\mathbf{Y} - \mathrm{E}\mathbf{Y})\}^\top \\ &\quad = [\mathrm{diag}(\mathbf{\Sigma}_{\mathbf{XX}})^{-1/2}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}]^\top\{\mathrm{diag}(\mathbf{\Sigma}_{\mathbf{XX}})^{1/2}\mathbf{B}^*\mathrm{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}\} \\ &\qquad\qquad + [\mathrm{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}\{\boldsymbol{\varepsilon} - \mathrm{E}(\boldsymbol{\varepsilon})\}]^\top,\end{aligned}$$

where diag denotes the diagonal matrix with diagonal elements identical to those of the argument. Obviously, all components of $\text{diag}(\mathbf{\Sigma}_{\mathbf{XX}})^{-1/2}\{\mathbf{X} - \text{E}(\mathbf{X})\}$ and $\text{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}\{\mathbf{Y} - \text{E}(\mathbf{Y})\}$ have mean 0 and variance 1. Then, by making the simultaneous substitutions,

$$\text{diag}(\mathbf{\Sigma}_{\mathbf{XX}})^{-1/2}\{\mathbf{X} - \text{E}(\mathbf{X})\} \mapsto \mathbf{X}_{\text{new}}, \quad \text{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}\{\mathbf{Y} - \text{E}(\mathbf{Y})\} \mapsto \mathbf{Y}_{\text{new}},$$

$$\text{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2}\{\boldsymbol{\varepsilon} - \text{E}(\boldsymbol{\varepsilon})\} \mapsto \boldsymbol{\varepsilon}_{\text{new}}, \quad \text{diag}(\mathbf{\Sigma}_{\mathbf{XX}})^{1/2}\mathbf{B}^*\text{diag}(\mathbf{\Sigma}_{\mathbf{YY}})^{-1/2} \mapsto \mathbf{B}_{\text{new}},$$

we can convert the original Model (1.1) into a new form where all components of the covariates and the responses have mean 0 and variance 1. In Model (1.2), the substitutions performed on $f(\mathbf{X})$ and $g(\mathbf{Y})$ can be absorbed into the functions f and g, respectively.

Now we briefly discuss the prediction problem for Model (1.2) under the elliptical copula multivariate response regression model, where given a value $\mathbf{X}^*$ of the covariate, we would like to predict the response $\mathbf{Y}^*$. Our discussion essentially follows Section 2.1 in [11], which deals with a variant of the Gaussian copula regression model. In the ideal setting where the transformation functions f and g and the coefficient matrix $\mathbf{B}^*$ are known, the oracle predictor for the median of $\mathbf{Y}^*$ is

$$\mathbf{U}^* = g^{-1}\{\mathbf{B}^{*\top} f(\mathbf{X}^*)\}.$$

This results from adapting Eq. (4) in [11] to our context using the conditionals of elliptical distributions; see, e.g., Theorem 2.18 in [12]. The oracle predictor $\mathbf{U}^*$ has a unique value, irrespective of the identifiability conditions. Prediction at other quantile levels of $\mathbf{Y}^*$ is also possible but more involved.

In practice, f, g and $\mathbf{B}^*$ are not available, and it is natural to substitute them by their estimates $\widehat{f}$, $\widehat{g}$ and $\widehat{\mathbf{B}}$ when the latter are available. Then, an empirical predictor for the median is given by

$$\widehat{\mathbf{U}}^* = \widehat{g}^{-1}\{\widehat{\mathbf{B}}^\top \widehat{f}(\mathbf{X}^*)\}, \tag{E.1}$$

where $\widehat{g}^{-1}$ is a suitable inverse of $\widehat{g}$. Therefore, accurate estimations of f, g and $\mathbf{B}^*$ are all important for achieving good prediction performance. We have addressed the estimation of $\mathbf{B}^*$ in this paper.

In the most commonly encountered Gaussian copula regression model, the estimators of f and g are also straightforward to construct. We focus our discussion on f because the treatment for g is analogous. In this case, because $f(\mathbf{X})$ is multivariate normal and all its components have mean 0 and variance 1 by our identification conditions, $f(x_1, \ldots, x_p) = (f_1(x_1), \ldots, f_p(x_p))^\top$ is explicitly given by $f_r = \Phi^{-1} \circ F_r$, where F_r is the marginal distribution function of the rth coordinate of $\mathbf{X}$, and Φ^{-1} is the $\mathcal{N}(0,1)$ quantile function. To obtain an estimator $\widehat{f}_r$ of f_r, it is typical to substitute F_r in $f_r = \Phi^{-1} \circ F_r$ by (sometimes a Winsorized, i.e., truncated version of) the empirical marginal distribution function $F^*_{n,k}$ in (F.1); see, e.g., Eq. (3.26) in [19] and Section 4 in [25]. In the more general case where $f(\mathbf{X})$ is elliptical, Φ^{-1} should be replaced by proper quantile functions for the marginals of $f(\mathbf{X})$ to obtain f.

Appendix F: Use of the normal-score rank correlation estimator in the Gaussian copula case

In this Appendix, we examine the properties of the normal-score rank correlation estimator $\mathbf{\Sigma}_n$, also known as van der Waerden correlation matrix, as an estimator of the copula correlation matrix when the underlying copula is Gaussian. The discussion is carried out in a general context. For space consideration, only a brief discussion on the application of $\mathbf{\Sigma}_n$ to the Gaussian copula regression problem is included at the end.

Let $\mathbf{X} \in \mathbb{R}^p$ be a continuous random vector with Gaussian copula and copula correlation matrix $\mathbf{\Sigma}$. Let $F_1, \ldots, F_p$ denote the marginal distribution functions of $\mathbf{X} = (X_1, \ldots, X_p)^\top$. For each $i \in [n]$, let $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,p})^\top$ be an independent copy of $\mathbf{X}$. We define the (rescaled version of the) empirical marginal distribution function for the kth coordinate of $\mathbf{X}$ as

$$F^*_{n,k}(t) = \frac{1}{n+1} \sum_{i \in [n]} \mathbf{1}(X_{i,k} \leq t) = \frac{n}{n+1} F_{n,k}(t). \tag{F.1}$$

Then, the normal-score rank correlation estimator or van der Waerden correlation matrix $\mathbf{\Sigma}_n = [r_{n,kk'}]_{k,k' \in [p]}$ of $\mathbf{\Sigma} = [r_{kk'}]_{k,k' \in [p]}$ is defined, for all $k, k' \in [p]$, as

$$r_{n,kk'} = \frac{\phi_n}{n} \sum_{i \in [n]} \Phi^{-1}\{F^*_{n,k}(X_{i,k})\} \times \Phi^{-1}\{F^*_{n,k'}(X_{i,k'})\}.$$

See, e.g., Eq. (7) on p. 113 in [15]. Here ϕ_n is a deterministic correction factor given by

$$\phi_n = \left[\frac{1}{n} \sum_{i \in [n]} \left\{\Phi^{-1}\left(\frac{i}{n+1}\right)\right\}^2\right]^{-1} = 1 + \mathcal{O}\{n^{-1}\ln(n)\}. \tag{F.2}$$

For each $i \in [n]$, define the Gaussianized observation

$$\mathbf{Z}_i^{(n)} \equiv (\Phi^{-1}\{F^*_{n,1}(E_{i,1})\}, \ldots, \Phi^{-1}\{F^*_{n,p}(E_{i,p})\})^\top \tag{F.3}$$

and let $\mathbf{\Sigma}_n$ be the corresponding sample covariance matrix, viz.

$$\mathbf{\Sigma}_n = \frac{\phi_n}{n} \sum_{i \in [n]} \mathbf{Z}_i^{(n)} \mathbf{Z}_i^{(n)\top}. \tag{F.4}$$

The correction ϕ_n is asymptotically negligible, but with it the diagonal elements of $\mathbf{\Sigma}_n$ all equal to 1, and so $\mathbf{\Sigma}_n$ becomes a genuine correlation matrix. The elements of $\mathbf{\Sigma}_n$ belong to multivariate rank order statistics that are common in the literature; see [15, 36] for some early references.

In fixed dimension, the van der Waerden correlation matrix $\mathbf{\Sigma}_n$ enjoys one crucial property: it is asymptotically semiparametrically efficient. More precisely, $\mathbf{\Sigma}_n$ achieves the asymptotic covariance matrix lower bound in the Hájek–Le Cam convolution theorem; see [19]. The analogous estimators based on Kendall's tau and Spearman's rho do not share this property.

Whether or not the dimension is fixed, it also stems readily from (F.4) that $\mathbf{\Sigma}_n$ is always positive semidefinite. As a result, when the Lasso estimator studied earlier in this paper is equipped with the estimator $\mathbf{\Sigma}_n$, it no longer requires the projection trick in (2.4) that sometimes leads to complication in the analysis, as in Section 3.3 or 5.3.1. Therefore, at least in fixed dimension, when estimating the copula correlation matrix for Gaussian copulas, the van der Waerden correlation matrix $\mathbf{\Sigma}_n$ should clearly be favored over the estimators based on Kendall's tau and Spearman's rho.

That the estimator $\mathbf{\Sigma}_n$ is semiparametrically efficient in the fixed-dimensional setting is equivalent to the characterization that $\mathbf{\Sigma}_n$ is asymptotically linear in the efficient influence function; see, e.g., Lemma 25.23 in [39]. The main goal of this Appendix is to establish a high-dimensional counterpart to the above statement in Theorem F.1. The theorem states that, in high dimensions, each element of the estimator $\mathbf{\Sigma}_n$ is the summation of a term which is linear in the efficient influence function (thus this term is asymptotically normal after scaling by $n^{1/2}$) and a remainder term which is, with high probability, uniformly small (at a rate n^{-1} up to logarithm factors) over all coordinates.

While the elements of the Kendall's tau or the Spearman's rho matrix are also asymptotically linear in their own influence function, the latter is inefficient. Thus, our result implies that the covariance matrix of $\mathbf{\Sigma}_n$ is smaller than both the Kendall's tau and the Spearman's rho estimators, so long as the remainder terms do not appreciably affect the covariance matrix. Further discussion is provided in Section F.3.

To state Theorem F.1, first define the efficient influence function ℓ for estimating elements of $\mathbf{\Sigma}$ as in, e.g., Theorem 3.1 in [19]. For any given correlation coefficient $\rho \in \mathbb{R}$ and arbitrary marginal distribution functions G and H, let

$$\begin{aligned}\ell(y,z;\rho,G,H) = \Phi^{-1}\{G(y)\}\Phi^{-1}\{H(z)\} - (\rho/2)[\{\Phi^{-1}(G(y))\}^2 \\ + \{\Phi^{-1}(H(z))\}^2] : (y,z) \in \mathbb{R}^2 \to \mathbb{R}.\end{aligned}$$

In the statement of Theorem F.1 and in its proof, detailed in Section F.2, C denotes a finite absolute constant that does not depend on n and p (nor on k, k', m, r which occur later). However, this constant C may change at each occurrence. Furthermore, $\lesssim$ denotes an inequality that holds up to such a C as the multiplicative factor.

Theorem F.1. *Assume that* $\ln(p) = o(n)$ *and* $\ln(n) = \mathcal{O}(p)$. *The normal-score rank correlation estimator* $\mathbf{\Sigma}_n = [r_{n,kk'}]_{k,k'\in[p]}$ *of* $\mathbf{\Sigma} = [r_{kk'}]_{k,k'\in[p]}$ *then satisfies*

$$r_{n,kk'} - r_{kk'} = \frac{1}{n}\sum_{i\in[n]} \ell(X_{i,k}, X_{i,k'}; r_{kk'}, F_k, F_{k'}) + R_{n,kk'}, \tag{F.5}$$

where the remainder term $R_{n,kk'}$ *satisfies, with probability at least* $1 - C/p^2$,

$$\max_{k,k'\in[p]} |R_{n,kk'}| \le C\{\ln(p)\ln^2(n) + \ln^{3/2}(p)\}/n. \tag{F.6}$$

The condition $\ln(p) = o(n)$ is common in the literature on high-dimensional statistics. The other condition $\ln(n) = \mathcal{O}(p)$ is purely for convenience: with it, $\delta_{n,p}$ in (F.7) admits the simple upper bound $C \ln^{1/2}(p)$. This simplification will be used repetitively without further mention.

F.1. Preliminaries

At a high level, the proof follows that of Theorem 3.1 in [19] up to Eq. (3.35). However, the derivation there is for the fixed-dimensional setting, and even then only heuristic, and with the remainder term simply stated as having order $o_p(n^{-1/2})$ in contrast to our (F.6). Theorem 3.1 in [19] was formally established using the result from [36], which in turn is strictly for the fixed-dimensional setting. The core of our proof is showing that the term $B_{21,kk'} - \overline{B}_{21,kk'}$ is small, which relies heavily on the realization that this term has a hidden canonical U-statistic structure.

To start, we need the following classical but important lemma for the convergence of the empirical distribution function under a weighted metric, which is Corollary 1 in Section 11.2 in [38]. We define

$$\delta_{n,p} = \ln^{1/2}\{p \ln(n)\}. \tag{F.7}$$

Lemma F.2. *Assume the same conditions as in Theorem F.1. Then for all $k \in [p]$ and $n \in \mathbb{N}$, there exist events $E_{4,k,n}$, each with probability at least $1 - 1/p^3$, such that for the same absolute constant $M_1 > 0$, we have*

$$\sup_{t\in\Lambda_{n,1}} |(F^*_{n,k} - F_k)(t)|/\sqrt{F_k(t) \wedge (1 - F_k)(t)} \le M_1\delta_{n,p}/\sqrt{n} \tag{F.8}$$

on $E_{4,k,n}$, where in general, $a \wedge b = \min(a, b)$,

$$\Lambda_{n,1} = \{t : (2\,M_1/3)^2\delta^2_{n,p}/n \le F_k(t) \le 1 - (2\,M_1/3)^2\delta^2_{n,p}/n\}$$

and $\delta_{n,p}$ is defined in (F.7). Thus, setting

$$\Lambda_{n,2} = \{t : 4\,M_1^2\delta^2_{n,p}/n \le F_k(t) \le 1 - 4\,M_1^2\delta^2_{n,p}/n\},$$

we have, on the event $E_{4,k,n}$,

$$\sup_{t\in\Lambda_{n,2}} |(F^*_{n,k} - F_k)(t)|/\{F_k(t) \wedge (1 - F_k)(t)\} \le 1/2. \tag{F.9}$$

Proof. For the quantities in Corollary 1 in Section 11.2 in [38], we choose

$$\lambda = M_1\delta_{n,p}, \quad a = (2/3)^2\lambda^2/n \to 0, \quad b = \delta = 1/2.$$

As $\lambda = 3\,\delta\sqrt{an}$, we can choose $\gamma^+ = 1 - \delta = 1/2$. We can also choose $\gamma^- = 1$. That (F.8) holds with probability at least $1-1/p^3$ then follows from the corollary, and the simple fact that $\|F^*_{n,k} - F_{n,k}\|_\infty \le 1/(n + 1)$, for large enough $M_1 > 0$. Then, (F.9) follows as a simple consequence. □

We also record two auxiliary lemmas whose proofs we omit. Below, ϕ denotes the $\mathcal{N}(0,1)$ density.

Lemma F.3. *We have* $\mathrm{d}\Phi^{-1}(u)/\mathrm{d}u = 1/\phi\{\Phi^{-1}(u)\}$ *and* $\mathrm{d}^2\Phi^{-1}(u)/\mathrm{d}^2u = \Phi^{-1}(u)/\phi^2\{\Phi^{-1}(u)\}$.

Lemma F.4. *For some absolute constant* $M_2 < \infty$,

$$\sup_{u\in(0,1)} \frac{u \wedge (1-u)}{\phi\{\Phi^{-1}(u)\}} \le M_2, \qquad \sup_{u\in(0,1)} \frac{|\Phi^{-1}(u)|}{\sqrt{2\ln[1/[2\{u\wedge(1-u)\}]]}} \le 1. \tag{F.10}$$

F.2. Proof of Theorem F.1

First, because the diagonal elements of $\mathbf{\Sigma}_n$ all equal to 1, $R_{n,kk'} = 0$ when $k = k'$ and so (F.6) clearly holds. Thus we focus on the case $k \neq k'$. We have the decomposition

$$\sqrt{n}\,(r_{n,kk'} - r_{kk'}) = \Xi_{1,n,kk'} + \cdots + \Xi_{4,n,kk'} + \mathcal{O}\{n^{-1/2}\ln(n)\}, \tag{F.11}$$

where

$$\Xi_{1,n,kk'} = \frac{1}{\sqrt{n}} \sum_{i\in[n]} [\Phi^{-1}\{F_k(X_{i,k})\} \times \Phi^{-1}\{F_{k'}(X_{i,k'})\} - r_{kk'}],$$

$$\Xi_{2,n,kk'} = \frac{1}{\sqrt{n}} \sum_{i\in[n]} [\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}] \times \Phi^{-1}\{F_{k'}(X_{i,k'})\},$$

$$\Xi_{3,n,kk'} = \frac{1}{\sqrt{n}} \sum_{i\in[n]} \Phi^{-1}\{F_k(X_{i,k})\} \times [\Phi^{-1}\{F^*_{n,k'}(X_{i,k'})\} - \Phi^{-1}\{F_{k'}(X_{i,k'})\}]$$

$$\Xi_{4,n,kk'} = \frac{1}{\sqrt{n}} \sum_{i\in[n]} [\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}], \\ \times [\Phi^{-1}\{F^*_{n,k'}(X_{i,k'})\} - \Phi^{-1}\{F_{k'}(X_{i,k'})\}],$$

and the $\mathcal{O}\{\ln(n)\,n^{-1/2}\}$ term in (F.11) comes from the factor ϕ_n given in (F.2).

We treat the terms $\Xi_{2,n,kk'}, \Xi_{3,n,kk'}, \Xi_{4,n,kk'}$ on the right-hand side of (F.11) in sequence. We define $a_n = 4\,M_1^2\delta_{n,p}^2/n$, where M_1 is the absolute constant in Lemma F.2, and the sets $A_{1,n} = (0, a_n] \cup (1-a_n, 1)$ and $A_{2,n} = (a_n, 1-a_n]$ which form a partition of the interval $(0,1)$.

Proposition F.5. *For the term* $\Xi_{2,n,kk'}$ *in* (F.11), *we have*

$$\Xi_{2,n,kk'} = -\frac{r_{kk'}}{2}\frac{1}{\sqrt{n}} \sum_{i\in[n]} [\Phi^{-1}\{F_k(X_{i,k})\}^2 - 1] + R_{1,n,kk'},$$

where the remainder term $R_{1,n,kk'}$ *satisfies, with probability at least* $1 - C/p^2$,

$$\max_{k,k'\in[p]} |R_{1,n,kk'}| \le Cn^{-1/2}\{\ln(p)\ln^2(n) + \ln^{3/2}(p)\}.$$

Proof. We treat the "tail region" where $i \in [n]$ satisfies $F_k(X_{i,k}) \in A_{1,n}$ separately from the region where $i \in [n]$ satisfies $F_k(X_{i,k}) \in A_{2,n}$. We write

$$\Xi_{2,n,kk'} \equiv B_{1,kk'} + B_{2,kk'}, \tag{F.12}$$

where

$$B_{1,kk'} = \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{1,n}} [\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}] \times \Phi^{-1}\{F_{k'}(X_{i,k'})\}$$

and

$$B_{2,kk'} = \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{2,n}} [\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}] \times \Phi^{-1}\{F_{k'}(X_{i,k'})\}.$$

Part 1: Treatment of the term $B_{1,kk'}$. We write $B_{1,kk'} = B_{11,kk'} - B_{12,kk'}$, where

$$B_{11,kk'} = \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{1,n}} \Phi^{-1}\{F^*_{n,k}(X_{i,k})\} \times \Phi^{-1}\{F_{k'}(X_{i,k'})\}$$

and

$$B_{12,kk'} = \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{1,n}} \Phi^{-1}\{F_k(X_{i,k})\} \times \Phi^{-1}\{F_{k'}(X_{i,k'})\}.$$

We only analyze the term $B_{12,kk'}$ in detail; this will culminate in the bound (F.16). The term $B_{11,kk'}$ is bounded similarly (with a possibly different C). This follows by an argument similar to that for $B_{12,kk'}$, but simply using the bound $\max_{i \in [n]} |\Phi^{-1}\{F^*_{n,k}(X_{i,k})\}| \lesssim \ln^{1/2}(n)$, which follows from the facts that

$$\min\{F^*_{n,k}(X_{i,k}), 1 - F^*_{n,k}(X_{i,k}) : i \in [n]\} \geq 1/(n+1),$$

and the right-hand inequality in (F.10).

We first control the expectation of $B_{12,kk'}$. We let $F_{kk'}$ be the bivariate distribution function of $(X_k, X_{k'})$, and let $F_{n,kk'}$ be its empirical counterpart based on $\{(X_{i,k}, X_{i,k'}) : i \in [n]\}$. We can write

$$\begin{aligned} B_{12,kk'} = n^{1/2} \int & \mathbf{1}\{F_k(y) \in A_{1,n}\} \\ & \times \Phi^{-1}\{F_k(y)\}\Phi^{-1}\{F_{k'}(z)\} \mathrm{d}F_{n,kk'}(y,z) \end{aligned} \tag{F.13}$$

and so

$$\begin{aligned} |\mathrm{E}B_{12,kk'}| &\leq n^{1/2} \int \mathbf{1}\{F_k(y) \in A_{1,n}\} |\Phi^{-1}\{F_k(y)\}| \times |\Phi^{-1}\{F_{k'}(z)\}| \mathrm{d}F_{kk'}(y,z) \\ &\leq n^{1/2} \int \mathbf{1}\{F_k(y) \in A_{1,n}\} |\Phi^{-1}\{F_k(y)\}| \times \{|\Phi^{-1}\{F_k(y)\}| + 1\} \mathrm{d}F_k(y) \end{aligned}$$

$$\begin{aligned}&\lesssim n^{1/2}\int \mathbf{1}\{u\in(0,a_n]\}\{\ln(1/2u)+\ln^{1/2}(1/2u)\}\mathrm{d}u\\ &\lesssim n^{-1/2}\ln(p)\ln(n). \qquad\qquad (\text{F.14})\end{aligned}$$

The second step follows by integrating over z upon conditioning on y, using the fact that given $X_{1,k}=y$, the variable $\Phi^{-1}\{F_{k'}(X_{1,k'})\}$ is Gaussian with mean $r_{kk'}\Phi^{-1}\{F_k(y)\}$ and variance $1-r_{kk'}^2$, and calling on Lemma B.3. In the third step, we have invoked the right-hand inequality in (F.10).

Next, we study the concentration of $B_{12,kk'}$ around $\mathrm{E}(B_{12,kk'})$. We will apply Bernstein's inequality, slightly simplified here as Lemma B.4. To this end, we bound the rth absolute moment, where $r\geq 2$, of the integrand in (F.13). We obtain

$$\begin{aligned}\int \mathbf{1}\{F_k(y)\in A_{1,n}\}&|\Phi^{-1}\{F_k(y)\}\Phi^{-1}\{F_{k'}(z)\}|^r\mathrm{d}F_{kk'}(y,z)\\ &\lesssim \int \mathbf{1}\{F_k(y)\in A_{1,n}\}|\Phi^{-1}\{F_k(y)\}|^r\\ &\quad\times\{2^{r-1}|\Phi^{-1}\{F_k(y)\}|^r+2^{r-1}(r-1)!!\}\mathrm{d}F_k(y), \qquad (\text{F.15})\end{aligned}$$

where we have again invoked Lemma B.3. We first concentrate on the first term in the curly bracket above. Using the right-hand inequality of (F.10), we have

$$\begin{aligned}2^{r-1}\int \mathbf{1}\{F_k(y)\in A_{1,n}\}&|\Phi^{-1}\{F_k(y)\}|^{2r}\mathrm{d}F_k(y)\\ &\lesssim 2^{2r-1}\int \mathbf{1}\{u\in(0,a_n]\}\ln^r(1/2u)\mathrm{d}u\end{aligned}$$

and the right-hand side can be successively rewritten as follows:

$$\begin{aligned}(-1)^r2^{2r-1}&\int \mathbf{1}\{u\in(0,a_n]\}\ln^r(2u)\mathrm{d}u\\ &=(-1)^r2^{2r-1}u\sum_{m=0}^{r}(-1)^{r-m}\frac{r!}{m!}\ln^m(2u)|_{u=a_n}\\ &=2^{2r-1}a_n\sum_{m=0}^{r}(-1)^m\frac{r!}{m!}\ln^m(2a_n).\end{aligned}$$

Furthermore, the right-hand most term in the above expression reduces to

$$r!\,2^{2r-1}a_n\sum_{m=0}^{r}\frac{1}{m!}\ln^m\{1/(2a_n)\}\leq r!\,2^{2r-1}a_n\sum_{m=0}^{r}\ln^m(1/2a_n).$$

Therefore,

$$2^{r-1}\int \mathbf{1}\{F_k(y)\in A_{1,n}\}|\Phi^{-1}\{F_k(y)\}|^{2r}\mathrm{d}F_k(y)\lesssim r!2^{2r-1}a_n\ln^r\{1/(2a_n)\}.$$

In the last step, we simply used the formula for the sum of a geometric series. For the second term in the curly bracket in (F.15), Hölder's inequality implies

$$
\begin{aligned}
&2^{r-1}(r-1)!! \int \mathbf{1}\{F_k(y) \in A_{1,n}\} |\Phi^{-1}\{F_k(y)\}|^r \mathrm{d}F_k(y) \\
&\lesssim 2^{3r/2-1}(r-1)!! \left\{ \int \mathbf{1}\{u \in (0, a_n]\} \ln^r(1/2u)\mathrm{d}u \right\}^{1/2} \left\{ \int \mathbf{1}\{u \in (0, a_n]\}\mathrm{d}u \right\}^{1/2} \\
&\lesssim 2^{3r/2-1}(r-1)!![r!a_n \ln^r\{1/(2a_n)\}]^{1/2} a_n^{1/2} \leq 2^{3r/2-1} r! a_n \ln^{r/2}\{1/(2a_n)\}.
\end{aligned}
$$

Therefore, overall the left-hand side of (F.15) can be bounded above by the multiple of a constant which is uniform over all integers $r \geq 2$, and the quantity

$$
r!2^{2r} a_n \ln^r\{1/(2a_n)\} = r![16 a_n \ln^2\{1/(2a_n)\}] \times [4 \ln\{1/(2a_n)\}]^{r-2}
$$

uniformly over r and n. We can then apply Bernstein's inequality with $c \lesssim \ln\{1/(2a_n)\}$ and $\sigma_i^2 \lesssim a_n \ln^2(1/2a_n)$, and conclude that

$$
\Pr[|B_{12,kk'} - \mathrm{E}(B_{12,kk'})| \geq Cn^{-1/2}\{\delta_{n,p} \ln(n)\sqrt{u} + \ln(n)u\}] \leq 2e^{-u}.
$$

Combining the above with the expectation bound (F.14) earlier, and setting $u = C\ln(p)$ for C large enough, we conclude that

$$
\Pr\{|B_{12,kk'}| \geq Cn^{-1/2} \ln(p) \ln(n)\} \leq 1/(2p^4). \tag{F.16}
$$

Together with an identical bound for $B_{11,kk'}$ (as argued earlier), we conclude that there exists an event $E_{5,kk',n}$ with probability at least $1 - 1/p^4$ on which

$$
|B_{1,kk'}| \leq Cn^{-1/2} \ln(p) \ln(n). \tag{F.17}
$$

Part 2: Treatment of the term $B_{2,kk'}$. By Taylor expansion to second order and Lemma F.3, the term $B_{2,kk'}$ can be written as

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{1}{\phi[\Phi^{-1}\{F_k(X_{i,k})\}]} (F_{n,k}^* - F_k)(X_{i,k}) \Phi^{-1}\{F_{k'}(X_{i,k'})\} \\
&\quad + \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{2,n}} \left[\frac{\Phi^{-1}(\overline{F}_{n,k,i}^*)}{\phi^2\{\Phi^{-1}(\overline{F}_{n,k,i}^*)\}} \right] (F_{n,k}^* - F_k)^2(X_{i,k}) \Phi^{-1}\{F_{k'}(X_{i,k'})\} \\
&\quad \equiv B_{21,kk'} + B_{22,kk'}.
\end{aligned} \tag{F.18}
$$

In the above, for each i, the quantity $\overline{F}_{n,k,i}^*$ is a random number strictly between $F_{n,k}^*(X_{i,k})$ and $F_k(X_{i,k})$. We can write $B_{21,kk'}$ equivalently as

$$
B_{21,kk'} = \int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{\Phi^{-1}\{F_{k'}(z)\}}{\phi[\Phi^{-1}\{F_k(y)\}]} \sqrt{n}\, (F_{n,k}^* - F_k)(y) \mathrm{d}F_{n,kk'}(y, z).
$$

We intend to approximate $B_{21,kk'}$ by

$$\overline{B}_{21,kk'} \equiv \int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{\Phi^{-1}\{F_{k'}(z)\}}{\phi[\Phi^{-1}\{F_k(y)\}]} \sqrt{n}\,(F_{n,k} - F_k)(y)\mathrm{d}F_{kk'}(y,z).$$

We now show that $B_{21,kk'} - \overline{B}_{21,kk'}$ is indeed small through a U-statistic argument. Introduce the functions $f_{n,kk'}, g_{n,kk'} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ specifically as

$$f_{n,kk'}(y_1,z_1;y_2,z_2) = \mathbf{1}\{F_k(y_1) \in A_{2,n}\} \frac{\Phi^{-1}\{F_{k'}(z_1)\}}{\phi[\Phi^{-1}\{F_k(y_1)\}]} \{\mathbf{1}(y_2 \le y_1) - F_k(y_1)\},$$

$$g_{n,kk'}(y_1,z_1;y_2,z_2) = f_{n,kk'}(y_1,z_1;y_2,z_2) - \int f_{n,kk'}(y,z;y_2,z_2)\mathrm{d}F_{kk'}(y,z),$$

and the index set $I_n^2 = \{(i,j) : i,j \in [n], i \neq j\}$. First replace $F_{n,k}^*$ by $F_{n,k}$ and then write out $F_{n,k}$ explicitly as in (F.1). We find

$$B_{21,kk'} - \overline{B}_{21,kk'} \equiv \Delta_{21,kk',1} + \Delta_{21,kk',2} + \Delta_{21,kk',3}, \tag{F.19}$$

where we have introduced the quantities

$$\Delta_{21,kk',1} = n^{-3/2} \sum_{(i,j)\in I_n^2} g_{n,kk'}(X_{i,k}, X_{i,k'}; X_{j,k}, X_{j,k'}),$$
$$\Delta_{21,kk',2} = n^{-3/2} \sum_{i\in[n]} g_{n,kk'}(X_{i,k}, X_{i,k'}; X_{i,k}, X_{i,k'}),$$
$$\Delta_{21,kk',3} = \frac{-\sqrt{n}}{n+1} \int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{\Phi^{-1}\{F_{k'}(z)\}}{\phi[\Phi^{-1}\{F_k(y)\}]} F_{n,k}(y)\mathrm{d}F_{n,kk'}(y,z).$$

The quantity $\Delta_{21,kk',1}$ is clearly (the "off-diagonal" part of) a U-statistic, with kernel $g_{n,kk'}$. Moreover,

$$\mathrm{E}\{g_{n,kk'}(X_k, X_{k'}; y_2, z_2)\} = 0, \quad \mathrm{E}\{g_{n,kk'}(y_1, z_1; X_k, X_{k'})\} = 0. \tag{F.20}$$

Indeed, the first equality follows by integration with respect to the measure $F_{kk'}$ whereas the second stems from the fact that $\mathrm{E}\{\mathbf{1}(X_k \le y_1) - F_k(y_1)\} = 0$.

Thus $g_{n,kk'}$ is canonical (i.e., completely degenerate) for the measure $F_{kk'}$. If the kernel $g_{n,kk'}$ were "nice," this would suggest that $\Delta_{21,kk',1}$ could be on the order of $n^{-1/2}$. However, this kernel is unbounded, and it diverges ever more quickly as n increases. Thus, establishing a deviation inequality for $\Delta_{21,kk',1}$ is quite tedious. The formal treatment of $\Delta_{21,kk',1}$, as well as the "leftover" terms $\Delta_{21,kk',2}$ (which is the "diagonal", i.e., the $i = j$ part of a U-statistic) and $\Delta_{21,kk',3}$, will be considered next.

Proposition F.6. *The terms $\Delta_{21,kk',1}$, $\Delta_{21,kk',2}$, $\Delta_{21,kk',3}$ satisfy*

$$\Pr[|\Delta_{21,kk',1}| > C\{\ln(p)\ln(n) + \ln^{3/2}(p)\}n^{-1/2}] \le C/p^4, \tag{F.21}$$
$$\Pr\{|\Delta_{21,kk',2}| + |\Delta_{21,kk',3}| \ge C\ln^{3/2}(n)n^{-1/2}\} \le 2/p^4. \tag{F.22}$$

Therefore, there exists an event $E_{6,kk',n}$ with probability at least $1 - C/p^4$ on which

$$|B_{21,kk'} - \overline{B}_{21,kk'}| \leq C\{\ln(p)\ln(n) + \ln^{3/2}(p) + \ln^{3/2}(n)\}n^{-1/2}. \qquad \text{(F.23)}$$

Proof. Let $h_{n,kk'}$ be the symmetrized version of $g_{n,kk'}(\cdot_{11}, \cdot_{12}; \cdot_{21}, \cdot_{22})$, i.e.,

$$\begin{aligned} h_{n,kk'}&(y_1, z_1; y_2, z_2) \\ &= \frac{1}{2}\{f_{n,kk'}(y_1, z_1; y_2, z_2) - \int f_{n,kk'}(y, z; y_2, z_2)\mathrm{d}F_{kk'}(y, z) \\ &\qquad + f_{n,kk'}(y_2, z_2; y_1, z_1) - \int f_{n,kk'}(y, z; y_1, z_1)\mathrm{d}F_{kk'}(y, z)\} \\ &\equiv (h_{n,kk',1} + h_{n,kk',2} + h_{n,kk',3} + h_{n,kk',4})(y_1, z_1; y_2, z_2)/2. \qquad \text{(F.24)} \end{aligned}$$

From (F.20), the variant with $g_{n,kk'}$ replaced by $h_{n,kk'}$ holds, too; thus $h_{n,kk'}$ is also canonical for the measure $F_{kk'}$.

We consider the "decoupled" version of the U-statistic $\Delta_{21,kk',1}$. For each $i \in [n]$, let $\mathbf{X}_i^* = (X_{i,1}^*, \ldots, X_{i,p}^*)$ be an independent copy of $\mathbf{X}_i$. The decoupled version of $\Delta_{21,kk',1}$, including the diagonal part to be precise, is then

$$\sum_{i,j\in[n]} g_{n,kk'}(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*) = \sum_{i,j\in[n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*).$$

We wish to apply Theorem 3.2 in [14] to bound, for every integer $m \geq 3$, the mth moment of the quantity above. Via Markov's inequality, this will lead to a tail probability bound for the quantity above and, via the decoupling theorem, to a tail probability bound for the un-decoupled $\Delta_{21,kk',1}$.

Let $\|\cdot\|_{L^2\to L^2}$ be defined as in (3.9) in [14]. Also for a (generic) random variable X, let E_X denote the expectation taken with respect to the random variable X only. Then, Theorem 3.2 in [14] states that there exists a universal constant $K < \infty$ such that (for the canonical kernel $h_{n,kk'}$ of two variables, and independent random variables $\mathbf{X}_i$, $\mathbf{X}_i^*$ for all $i \in [n]$), for every integer $m \geq 2$,

$$\begin{aligned} &\mathrm{E}\Big|\sum_{i,j\in[n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*)\Big|^m \\ &\leq K^m\Bigg[m^{m/2}\Big[\sum_{i,j\in[n]} \mathrm{E}\{h_{n,kk'}^2(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*)\}\Big]^{m/2} + m^m\|h_{n,kk'}\|_{L^2\to L^2}^m \\ &\quad + m^{3m/2}\mathrm{E}_{\{\mathbf{X}_i, i\in[n]\}} \max_{i\in[n]}\Big[\sum_{j\in[n]} \mathrm{E}_{\mathbf{X}_j^*}\{h_{n,kk'}^2(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*)\}\Big]^{m/2} \\ &\quad + m^{2m}\mathrm{E}\Big\{\max_{i,j\in[n]} |h_{n,kk'}^m(X_{i,k}, X_{i,k'}; X_{j,k}^*, X_{j,k'}^*)|\Big\}\Bigg]. \qquad \text{(F.25)} \end{aligned}$$

Note that while the statement of the theorem requires a bounded kernel, inspection of the proof reveals that it is not necessary. If in doubt, we can always truncate the unbounded kernel $h_{n,kk'}$, and then let the truncation level go to infinity on both sides of (F.25).

In what follows, sometimes we will write $h_{n,kk'}$ simply as h. To apply (F.25), we need to bound four quantities, namely

(i) $\mathrm{E}\{h^2(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\}$;
(ii) $\|h\|_{L^2\to L^2}$;
(iii) $\mathrm{E}_{\{\mathbf{X}_i, i\in[n]\}}\Big[\max_{i\in[n]}\{\mathrm{E}_{\mathbf{X}^*_j} h^2(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\}^{m/2}\Big]$;
(iv) $\mathrm{E}\max_{i,j\in[n]} |h^m(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})|$.

We carry out these tasks in sequence. For (i), first by Jensen's inequality,

$$\mathrm{E}\{h^2(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\} \le 4\,\mathrm{E}\{f^2_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\}$$

and the latter can be rewritten as follows:

$$\begin{aligned}
&4\,\mathrm{E}\Bigg[\mathbf{1}\{F_k(X_{i,k}) \in A_{2,n}\}\left\{\frac{\Phi^{-1}\{F_{k'}(X_{i,k'})\}}{\phi[\Phi^{-1}\{F_k(X_{i,k})\}]}\right\}^2 \\
&\qquad\qquad\qquad\qquad \times \{\mathbf{1}(X^*_{j,k} \le X_{i,k}) - F_k(X_{i,k})\}^2\Bigg] \\
&= 4\,\mathrm{E}_{\mathbf{X}_i}\Bigg[\mathbf{1}\{F_k(X_{i,k}) \in A_{2,n}\}\left\{\frac{\Phi^{-1}\{F_{k'}(X_{i,k'})\}}{\phi[\Phi^{-1}\{F_k(X_{i,k})\}]}\right\}^2 \\
&\qquad\qquad\qquad\qquad \times \mathrm{E}_{\mathbf{X}^*_j}\{\mathbf{1}(X^*_{j,k} \le X_{i,k}) - F_k(X_{i,k})\}^2\Bigg] \\
&= 4\,\mathrm{E}_{\mathbf{X}_i}\Bigg[\mathbf{1}\{F_k(X_{i,k}) \in A_{2,n}\}\left\{\frac{\Phi^{-1}\{F_{k'}(X_{i,k'})\}}{\phi[\Phi^{-1}\{F_k(X_{i,k})\}]}\right\}^2 \\
&\qquad\qquad\qquad\qquad \times F_k(X_{i,k})\,\{1 - F_k(X_{i,k})\}\Bigg].
\end{aligned}$$

Therefore, upon invoking Lemma B.3 and then (F.10), we can deduce that

$$\mathrm{E}\{h^2(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\} \le C\ln^2\{1/(2a_n)\}. \tag{F.26}$$

For (ii), first by definition (see (3.9) in [14]), we can write $\|h\|_{L^2\to L^2}$ as

$$\begin{aligned}
\sup\Bigg\{&\mathrm{E}\sum_{i,j\in[n]} h(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'}) f_i(X_{i,k}, X_{i,k'}) g_j(X^*_{j,k}, X^*_{j,k'}) : \\
&\qquad\qquad \mathrm{E}\sum_{i\in[n]} f_i^2(X_{i,k}, X_{i,k'}) \le 1, \mathrm{E}\sum_{j\in[n]} g_j^2(X^*_{j,k}, X^*_{j,k'}) \le 1\Bigg\}
\end{aligned}$$

so that, by Jensen's inequality,

$$\|h\|_{L^2\to L^2} \le \sup\Bigg\{ \sum_{i,j\in[n]} \left\{\mathrm{E}h^2(X_{i,k},X_{i,k'};X^*_{j,k},X^*_{j,k'})\right\}^{1/2} \times \{\mathrm{E}f_i^2(X_{i,k},X_{i,k'})\}^{1/2} \times \left\{\mathrm{E}g_j^2(X^*_{j,k},X^*_{j,k'})\right\}^{1/2} : \mathrm{E}\sum_{i\in[n]} f_i^2(X_{i,k},X_{i,k'}) \le 1, \mathrm{E}\sum_{j\in[n]} g_j^2(X^*_{j,k},X^*_{j,k'}) \le 1\Bigg\}.$$

Now, by (F.26), the right-hand side of this inequality can be bounded above by

$$C\ln\{1/(2a_n)\} \times \sup\left\{\left\{\sum_{i,j\in[n]} \mathrm{E}f_i^2(X_{i,k},X_{i,k'})\right\}^{1/2}\left\{\sum_{i,j\in[n]} \mathrm{E}g_j^2(X^*_{j,k},X^*_{j,k'})\right\}^{1/2} : \mathrm{E}\sum_{i\in[n]} f_i^2(X_{i,k},X_{i,k'}) \le 1, \mathrm{E}\sum_{j\in[n]} g_j^2(X^*_{j,k},X^*_{j,k'}) \le 1\right\},$$

and hence

$$\|h\|_{L^2\to L^2} \le Cn\ln\{1/(2a_n)\}. \tag{F.27}$$

For (iii), we first compute the inner expectation with respect to $\mathbf{X}_j^*$. To this end, the contributions from the squares of the four terms in the curly bracket in (F.24) are computed separately. For the first term involving $h_{n,kk',1}$,

$$\mathrm{E}_{\mathbf{X}_j^*}\{h^2_{n,kk',1}(y_1,z_1;X^*_{j,k},X^*_{j,k'})\} = \mathrm{E}_{\mathbf{X}_j^*}\{f^2_{n,kk'}(y_1,z_1;X^*_{j,k},X^*_{j,k'})\}.$$

Then, realizing that $\mathbf{1}(X^*_{j,k}\le y_1)$ is Bernoulli with success probability $F_k(y_1)$, we can rewrite the right-hand term as

$$\begin{aligned}\mathrm{E}_{\mathbf{X}_j^*}&\left[\mathbf{1}\{F_k(y_1)\in A_{2,n}\}\frac{\Phi^{-1}\{F_{k'}(z_1)\}}{\phi[\Phi^{-1}\{F_k(y_1)\}]}\{\mathbf{1}(X^*_{j,k}\le y_1)-F_k(y_1)\}\right]^2 \\ &= \mathbf{1}\{F_k(y_1)\in A_{2,n}\}\left[\frac{\Phi^{-1}\{F_{k'}(z_1)\}}{\phi[\Phi^{-1}\{F_k(y_1)\}]}\right]^2 F_k(y_1)\{1-F_k(y_1)\}.\end{aligned}$$

For the second term involving $h_{n,kk',2}$, given that

$$h_{n,kk',2}(y_1,z_1;y_2,z_2) = -\mathrm{E}_{\mathbf{X}_i}\{f_{n,kk'}(X_{i,k},X_{i,k'};y_2,z_2)\},$$

it follows from Jensen's inequality that

$$h^2_{n,kk',2}(y_1,z_1;y_2,z_2) \le \mathrm{E}_{\mathbf{X}_i}\{f^2_{n,kk'}(X_{i,k},X_{i,k'};y_2,z_2)\}$$

and hence

$$\mathrm{E}_{\mathbf{X}_j^*}\{h^2_{n,kk',2}(y_1,z_1;X^*_{j,k},X^*_{j,k'})\}$$

$$\le \mathrm{E}_{\mathbf{X}_j^*}\mathrm{E}_{\mathbf{X}_i}\{f_{n,kk'}^2(X_{i,k},X_{i,k'};X_{j,k}^*,X_{j,k'}^*)\}$$
$$= \mathrm{E}\{f_{n,kk'}^2(X_{i,k},X_{i,k'};X_{j,k}^*,X_{j,k'}^*)\} \le C\ln^2\{1/(2a_n)\},$$

where the last step follows from (F.26). For the third term involving $h_{n,kk',3}$, note that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_j^*}\{h_{n,kk',3}^2(y_1,z_1;X_{j,k}^*,X_{j,k'}^*)\}\\
&\quad= \mathrm{E}_{\mathbf{X}_j^*}\left[\mathbf{1}\{F_k(X_{j,k}^*)\in A_{2,n}\}\frac{\Phi^{-1}\{F_{k'}(X_{j,k'}^*)\}}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]}\{\mathbf{1}(y_1\le X_{j,k}^*)-F_k(X_{j,k}^*)\}\right]^2\\
&\quad= \mathrm{E}_{\mathbf{X}_j^*}\Bigg[\mathbf{1}\{F_k(X_{j,k}^*)\in A_{2,n}\}\frac{\Phi^{-1}\{F_{k'}(X_{j,k'}^*)\}}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]}\\
&\qquad\qquad\times[\mathbf{1}\{F_k(y_1)\le F_k(X_{j,k}^*)\}-F_k(X_{j,k}^*)]\Bigg]^2. \qquad \text{(F.28)}
\end{aligned}$$

To be precise, the last equality holds with probability 1 when eventually we set $y_1 = X_{i,k}$. If $F_k(y_1)\le a_n$, (F.28) can be rewritten and bounded from above as follows:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_j^*}\left[\mathbf{1}\{F_k(X_{j,k}^*)\in A_{2,n}\}\frac{\Phi^{-1}\{F_{k'}(X_{j,k'}^*)\}}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]}\left\{1-F_k(X_{j,k}^*)\right\}\right]^2\\
&\quad\le C\mathrm{E}_{\mathbf{X}_j^*}\left[\mathbf{1}\{F_k(X_{j,k}^*)\in A_{2,n}\}\frac{\Phi^{-1}\{F_k(X_{j,k}^*)\}^2+1}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]^2}\left\{1-F_k(X_{j,k}^*)\right\}^2\right]\\
&\qquad= C\int_{a_n}^{1-a_n}\frac{\Phi^{-1}(u)^2+1}{\phi\{\Phi^{-1}(u)\}^2}(1-u)^2\mathrm{d}u \le Ca_n^{-1}\ln\{1/(2a_n)\}.
\end{aligned}$$

If $a_n < F_k(y_1)\le 1/2$, dividing the range of integration based on the position of $F_k(X_{j,k}^*)$ relative to $F_k(y_1)$ and $1/2$, and making multiple uses of (F.10), we can rewrite (F.28) and bound it as follows:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_j^*}\left[\mathbf{1}\{F_k(X_{j,k}^*)\in(a_n,F_k(y_1))\}\frac{\Phi^{-1}\{F_{k'}(X_{j,k'}^*)\}}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]}\left\{-F_k(X_{j,k}^*)\right\}\right]^2\\
&\quad+\mathrm{E}_{\mathbf{X}_j^*}\left[\mathbf{1}\{F_k(X_{j,k}^*)\in[F_k(y_1),1-a_n]\}\frac{\Phi^{-1}\{F_{k'}(X_{j,k'}^*)\}}{\phi[\Phi^{-1}\{F_k(X_{j,k}^*)\}]}\left\{1-F_k(X_{j,k}^*)\right\}\right]^2\\
&\quad\le C\int_{a_n}^{F_k(y_1)}\frac{\Phi^{-1}(u)^2+1}{\phi\{\Phi^{-1}(u)\}^2}u^2\mathrm{d}u + C\int_{F_k(y_1)}^{1/2}\frac{\Phi^{-1}(u)^2+1}{\phi\{\Phi^{-1}(u)\}^2}(1-u)^2\,\mathrm{d}u\\
&\qquad\qquad+C\int_{1/2}^{1-a_n}\frac{\Phi^{-1}(u)^2+1}{\phi\{\Phi^{-1}(u)\}^2}(1-u)^2\,\mathrm{d}u\\
&\quad\le C\int_{a_n}^{F_k(y_1)}[\ln\{1/(2u)\}+1]\mathrm{d}u + C\int_{F_k(y_1)}^{1/2}[\ln\{1/(2u)\}+1]\,\frac{(1-u)^2}{u^2}\,\mathrm{d}u
\end{aligned}$$

$$+ C \int_{1/2}^{1-a_n} [\ln\{1/2(1-u)\} + 1] \,\mathrm{du}$$
$$\leq C \left[\frac{1}{F_k(y_1)} \ln[1/\{2F_k(y_1)\}] + F_k(y_1) \right].$$

As the case where $F_k(y_1) > 1/2$ is analogous, we conclude that

$$\mathrm{E}_{\mathbf{X}_j^*}\{h^2_{n,kk',3}(y_1, z_1; X^*_{j,k}, X^*_{j,k'})\}$$
$$\leq C \Bigg[\frac{1}{a_n \vee \{F_k \wedge (1-F_k)\}(y_1)} \ln \left[\frac{1}{2[a_n \vee \{F_k \wedge (1-F_k)\}(y_1)]} \right] + \{F_k \wedge (1-F_k)\}(y_1) \Bigg]. \quad \text{(F.29)}$$

For the fourth term involving $h_{n,kk',4}$, note that

$$h_{n,kk',4}(y_1, z_1; y_2, z_2) = -\mathrm{E}\{h_{n,kk',3}(y_1, z_1; X_{j,k}, X_{j,k'})\}.$$

Therefore, by Jensen's inequality,

$$\mathrm{E}_{\mathbf{X}_j^*}\{h^2_{n,kk',4}(y_1, z_1; X^*_{j,k}, X^*_{j,k'})\}$$
$$= \mathrm{E}_{\mathbf{X}_j^*}[\mathrm{E}\{h_{n,kk',3}(y_1, z_1; X_{j,k}, X_{j,k'})\}]^2$$
$$\leq \mathrm{E}\{h^2_{n,kk',3}(y_1, z_1; X_{j,k}, X_{j,k'})\},$$

which then admits the same bound as (F.29).

Thus, for the term (iii), we first have

$$\mathrm{E}_{\mathbf{X}_j^*}\{h^2(y_1, z_1; X^*_{j,k}, X^*_{j,k'})\} \leq \bar{h}(y_1, z_1),$$

where the right-hand side is defined by

$$\mathbf{1}\{F_k(y_1) \in A_{2,n}\} \left[\frac{\Phi^{-1}\{F_{k'}(z_1)\}}{\phi[\Phi^{-1}\{F_k(y_1)\}]} \right]^2 F_k(y_1)\{1 - F_k(y_1)\}$$
$$+ C \Bigg[\ln^2\{1/(2a_n)\} + \{F_k \wedge (1-F_k)\}(y_1) + \frac{1}{a_n \vee \{F_k \wedge (1-F_k)\}(y_1)} \ln \left[\frac{1}{2[a_n \vee \{F_k \wedge (1-F_k)\}(y_1)]} \right] \Bigg].$$

For $m \geq 3$, the expectation of $\bar{h}^{m/2}(X_{i,k}, X_{i,k'})$ can be bounded above as follows:

$$\mathrm{E}\{\bar{h}^{m/2}(X_{i,k}, X_{i,k'})\}$$
$$\leq C^m \int \Bigg[\mathbf{1}\{F_k(y) \in A_{2,n}\} \left[\frac{\Phi^{-1}\{F_{k'}(z)\}}{\phi[\Phi^{-1}\{F_k(y)\}]} \right]^2$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad \times F_k(y)\left\{1-F_k(y)\right\}\Bigg]^{m/2} \mathrm{d}F_{kk'}(y,z)\\
&+C^m \int \Bigg[\frac{1}{a_n \vee \{F_k \wedge (1-F_k)\}(y)}\\
&\qquad\qquad \ln\left[\frac{1}{2[a_n \vee \{F_k \wedge (1-F_k)\}(y)]}\right]\Bigg]^{m/2} \mathrm{d}F_k(y)\\
&+C^m \int \{F_k \wedge (1-F_k)\}^{m/2}(y)\mathrm{d}F_k(y) + C^m \ln^{2(m/2)}(1/2a_n)\\
&\le C^m \ln^{m/2}\{1/(2a_n)\} a_n^{1-m/2} + C^m m^{m/2} a_n^{1-m/2}.
\end{aligned}
$$

Then, simply bounding the maximum over all $i \in [n]$ by the summation over $i \in [n]$, we find

$$
\begin{aligned}
&\mathrm{E}_{\{\mathbf{X}_i, i\in[n]\}}\left[\max_{i\in[n]}\{\mathrm{E}_{\mathbf{X}_j^*} h^2(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\}^{m/2}\right]\\
&\qquad \le \mathrm{E}\Big\{\sum_{i\in[n]} \bar{h}(X_{i,k}, X_{i,k'})^{m/2}\Big\}\\
&\qquad \le C^m n[\ln^{m/2}\{1/(2a_n)\} a_n^{1-m/2} + m^{m/2} a_n^{1-m/2}]. \qquad \text{(F.30)}
\end{aligned}
$$

Finally, for the term (iv), first compute, for any integer $m \ge 3$,

$$
\mathrm{E}|h^m(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})| \le C^m[\ln^{m/2}\{1/(2a_n)\} a_n^{2-m} + m^{m/2} a_n^{2-m}].
$$

Then, bounding the maximum over $i, j \in [n]$ by the summation over $i, j \in [n]$, we obtain

$$
\begin{aligned}
&\mathrm{E} \max_{i,j\in[n]} |h^m(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})|\\
&\qquad\qquad \le C^m n^2[\ln^{m/2}\{1/(2a_n)\} a_n^{2-m} + m^{m/2} a_n^{2-m}]. \qquad \text{(F.31)}
\end{aligned}
$$

Now we are ready to apply (F.25). Combining (F.26), (F.27), (F.30), (F.31), we obtain

$$
\begin{aligned}
&\mathrm{E}\left|\sum_{i,j\in[n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})\right|^m\\
&\qquad \le C^m \max\{m^{m/2}(A_m)^m, m^m(B_m)^m, m^{3m/2}(C_m)^m,\\
&\qquad\qquad m^{2m}(D_m)^m, m^{5m/2}(E_m)^m\} \qquad \text{(F.32)}
\end{aligned}
$$

for the quantities

$$
\begin{aligned}
A_m = B_m = \ln\{1/(2a_n)\}\, n, \quad C_m = \ln^{1/2}\{1/(2a_n)\} a_n^{-1/2+1/m} n^{1/2+1/m},\\
D_m = a_n^{-1/2+1/m} n^{1/2+1/m} + \ln^{1/2}\{1/(2a_n)\} a_n^{-1+2/m} n^{2/m},
\end{aligned}
$$

and $E_m = a_n^{-1+2/m} n^{2/m}$. Now, let

$$t_m = Ce \max(m^{1/2} A_m, mB_m, m^{3/2} C_m, m^2 D_m, m^{5/2} E_m), \tag{F.33}$$

where the constant C could be the same as that in (F.32). By Markov's inequality, we have

$$\begin{aligned} &\Pr\Big(\Big| \sum_{i,j \in [n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'}) \Big| > t_m \Big) \\ &\qquad \le \frac{1}{t_m^m} \mathrm{E} \Big| \sum_{i,j \in [n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'}) \Big|^m \le e^{-m} \end{aligned} \tag{F.34}$$

for every integer $m \ge 3$. Now let $m = \lceil 4 \ln(p) \rceil$, so that the last line in (F.34) is $e^{-m} \le 1/p^4$. For this value of m, t_m in (F.33) admits the bound $t_m \le C\{\ln(p)\ln(n) + \ln^{3/2}(p)\}n$. Thus we conclude that

$$\begin{aligned} &\Pr\Big[\Big| \sum_{i,j \in [n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'}) \Big| \\ &\qquad > C\{\ln(p)\ln(n) + \ln^{3/2}(p)\}n \Big] \le 1/p^4. \end{aligned} \tag{F.35}$$

To properly apply the decoupling theorem, we also need to treat the "diagonal" component of $\sum_{i,j \in [n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{j,k}, X^*_{j,k'})$, namely (up to scaling by $n^{-3/2}$)

$$\Delta_{21,kk',4} \equiv n^{-3/2} \sum_{i \in [n]} h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{i,k}, X^*_{i,k'}).$$

It is easy to check that for each $i \in [n]$, $h_{n,kk'}(X_{i,k}, X_{i,k'}; X^*_{i,k}, X^*_{i,k'})$ is centered, and that for arbitrary integer $r \ge 2$, its rth absolute moment can be bounded above by a constant multiple of

$$\begin{aligned} &\int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{|\Phi^{-1}\{F_{k'}(z_1)\}|^r}{\phi[\Phi^{-1}\{F_k(y_1)\}]^r} \\ &\qquad\qquad \times |\mathbf{1}(y_2 \le y_1) - F_k(y_1)|^r \mathrm{d}F_{kk'}(y_1, z_1) \mathrm{d}F_{kk'}(y_2, z_2) \\ &\quad \le \int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{|\Phi^{-1}\{F_{k'}(z_1)\}|^r}{\phi[\Phi^{-1}\{F_k(y_1)\}]^r} \mathrm{d}F_{kk'}(y_1, z_1) \\ &\quad \le 2^{r-1} \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{|\Phi^{-1}\{F_k(y_1)\}|^r + (r-1)!!}{\phi[\Phi^{-1}\{F_k(y_1)\}]^r} \mathrm{d}F_k(y_1) \\ &\quad \lesssim 2^{3r/2-1} M_2^r \int \mathbf{1}\{u \in (a_n, 1/2]\} \frac{\ln^{r/2}(1/2u)}{u^r} \mathrm{d}u \\ &\qquad\qquad + 2^{r-1} M_2^r (r-1)!! \int \mathbf{1}\{u \in (a_n, 1/2]\} \frac{1}{u^r} \mathrm{d}u \\ &\quad \lesssim 2^{3r/2-1} M_2^r \ln^{r/2}\{1/(2a_n)\} a_n^{-r+1} + 2^{r-1} M_2^r (r-1)!! a_n^{-r+1} \end{aligned}$$

$$\lesssim r!\{\ln\{1/(2a_n)\}a_n^{-1}\}[2^{3/2}M_2\ln^{1/2}\{1/(2a_n)\}a_n^{-1}]^{r-2}.$$

Then, by Bernstein's inequality,

$$\Pr\{|\Delta_{21,kk',4}| \geq C\ln^{1/2}(n)n^{-1/2}\} \leq 1/p^4. \tag{F.36}$$

Next, first by (F.35), (F.36) and basic probability axioms, and then by decoupling using, e.g., Theorem 3.4.1 in [7], we can conclude that there exists an absolute constant $K \in (0,\infty)$ such that

$$\begin{aligned}2/p^4 &\geq \Pr\Big[\Big|\sum_{i,j\in[n]} h_{n,kk'}(X_{i,k},X_{i,k'};X^*_{j,k},X^*_{j,k'})\Big| \\ &\qquad > C\{\ln(p)\ln(n)+\ln^{3/2}(p)\}n\Big] + \Pr\{n^{3/2}|\Delta_{21,kk',4}| \geq C\ln^{1/2}(n)n\} \\ &\geq \Pr\Big[\Big|\sum_{(i,j)\in I_n^2} h_{n,kk'}(X_{i,k},X_{i,k'};X^*_{j,k},X^*_{j,k'})\Big| > C\{\ln(p)\ln(n)+\ln^{3/2}(p)\}n\Big] \\ &\geq \frac{1}{K}\Pr\Big[\Big|\sum_{(i,j)\in I_n^2} h_{n,kk'}(X_{i,k},X_{i,k'};X_{j,k},X_{j,k'})\Big| \\ &\qquad\qquad > KC\{\ln(p)\ln(n)+\ln^{3/2}(p)\}n\Big].\end{aligned}$$

Then, recalling that $\Delta_{21,kk',1}$ contains an overall $n^{-3/2}$ factor, (F.21) in the proposition follows.

We then turn to the term $\Delta_{21,kk',2}$. To treat it, write

$$\begin{aligned}\Delta_{21,kk',2} &= n^{-3/2}\sum_{i\in[n]} f_{n,kk'}(X_{i,k},X_{i,k'};X_{i,k},X_{i,k'}) \\ &\qquad - n^{-3/2}\sum_{i\in[n]} \mathrm{E}_{\mathbf{X}_i^*}\{f_{n,kk'}(X^*_{i,k},X^*_{i,k'};X_{i,k},X_{i,k'})\}.\end{aligned}$$

For the summand $f_{n,kk'}(X_{i,k},X_{i,k'};X_{i,k},X_{i,k'})$ in the first term on the right-hand side, we can bound the expectation of $|f_{n,kk'}(X_{i,k},X_{i,k'};X_{i,k},X_{i,k'})|$ from above by $C\ln^{3/2}\{1/(2a_n)\}$, while for any integer $r \geq 2$, its rth absolute moment can be bounded similarly as that for $\Delta_{21,kk',4}$.

As for the summand $\mathrm{E}_{\mathbf{X}_i^*}\{f_{n,kk'}(X^*_{i,k},X^*_{i,k'};X_{i,k},X_{i,k'})\}$ in the second term on the right-hand side above, its expectation is zero, while for any integer $r \geq 2$, its rth absolute moment can be bounded via Jensen's inequality as

$$\begin{aligned}&\mathrm{E}|\mathrm{E}_{\mathbf{X}_i^*}\{f_{n,kk'}(X^*_{i,k},X^*_{i,k'};X_{i,k},X_{i,k'})\}|^r \\ &\qquad\qquad \leq \mathrm{E}\{\mathrm{E}_{\mathbf{X}_i^*}|f_{n,kk'}(X^*_{i,k},X^*_{i,k'};X_{i,k},X_{i,k'})|^r\},\end{aligned}$$

which can again be bounded similarly as that for $\Delta_{21,kk',4}$. Thus,

$$\Pr\{|\Delta_{21,kk',2}| \geq C\ln^{3/2}(n)n^{-1/2}\} \leq 1/p^4.$$

Analogously, $\Delta_{21,kk',3}$ satisfies the same bound. Then, (F.22) in the proposition follows.

Finally, (F.23) holds as a simple consequence of (F.19), (F.21) and (F.22). This completes the proof of Proposition F.6. □

Next, we rewrite

$$\begin{aligned}\overline{B}_{21,kk'} &= r_{kk'} \int \mathbf{1}\{F_k(y) \in A_{2,n}\} \frac{\Phi^{-1}\{F_k(y)\}}{\phi[\Phi^{-1}\{F_k(y)\}]} \sqrt{n}\,(F_{n,k} - F_k)(y)\mathrm{d}F_k(y) \\ &= r_{kk'} \int \mathbf{1}\{F_k(y) \in A_{2,n}\}\Phi^{-1}\{F_k(y)\}\sqrt{n}\,(F_{n,k} - F_k)(y)\mathrm{d}\Phi^{-1}\{F_k(y)\} \\ &= \frac{r_{kk'}}{2}\left\{-\int \mathbf{1}\{F_k(y) \in A_{2,n}\}\sqrt{n}\,\Phi^{-1}\{F_k(y)\}^2\mathrm{d}(F_{n,k} - F_k)(y) + D_2\right\},\end{aligned}$$

so that

$$\overline{B}_{21,kk'} = -\frac{r_{kk'}}{2}\frac{1}{\sqrt{n}}\sum_{i\in[n]}[\Phi^{-1}\{F_k(X_{i,k})\}^2 - 1] + \frac{r_{kk'}D_1}{2} + \frac{r_{kk'}D_2}{2}. \tag{F.37}$$

The first step follows by integrating over z upon conditioning on y and the fact that given $X_k = y$, the variable $\Phi^{-1}\{F_{k'}(X_{k'})\}$ is Gaussian with mean $r_{kk'}\,\Phi^{-1}\{F_k(y)\}$. The third step and (F.37) follow by integration by parts, and upon setting

$$D_1 = \int \mathbf{1}\{F_k(y) \in A_{1,n}\}\sqrt{n}\,\Phi^{-1}\{F_k(y)\}^2\mathrm{d}(F_{n,k} - F_k)(y)$$

and

$$\begin{aligned}D_2 = \sqrt{n}\,\Phi^{-1}(1 - a_n)^2(F_{n,k} - F_k)\{F_k^{\leftarrow}(1 - a_n)\}& \\ - \sqrt{n}\,\Phi^{-1}(a_n)^2(F_{n,k} - F_k)\{F_k^{\leftarrow}(a_n)\}&.\end{aligned}$$

The term D_1 has expectation zero, and its concentration around expectation can be treated in a similar way as $B_{12,kk'}$ earlier to arrive at a bound similar to (F.16). Specifically, there exists an event $E_{7,k,n}$ with probability at least $1-1/p^3$ on which

$$|D_1| \le Cn^{-1/2}\ln(p)\ln(n). \tag{F.38}$$

Lemma F.2 also easily ensures that on $E_{4,k,n}$,

$$|D_2| \le Cn^{-1/2}\ln(p)\ln(n). \tag{F.39}$$

We finally treat the term $B_{22,kk'}$ from (F.18). By the mean value theorem and (F.9) on the event $E_{4,k,n}$, we have, on the same event, and for $F_k(y) \in A_{1,n}$,

$$|\Phi^{-1}(\overline{F}^*_{n,k,i}) - \Phi^{-1}\{F_k(X_{i,k})\}| = |\overline{F}^*_{n,k,i} - F_k(X_{i,k})|/\phi\{\Phi^{-1}(\widetilde{F}^*_{n,k,i})\}$$

$$\lesssim \frac{|F^*_{n,k}(X_{i,k}) - F_k(X_{i,k})|}{\widetilde{F}^*_{n,k,i} \wedge (1 - \widetilde{F}^*_{n,k,i})} \lesssim \frac{|F^*_{n,k}(X_{i,k}) - F_k(X_{i,k})|}{\{F_k \wedge (1 - F_k)\}(X_{i,k})} \leq \frac{1}{2}.$$

Here $\widetilde{F}^*_{n,k,i}$ is a random number strictly between $\overline{F}^*_{n,k,i}$ and $F_k(X_{i,k})$. Therefore, $|B_{22,kk'}|$ can be bounded above by

$$\begin{aligned}
&\sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{|\Phi^{-1}(\overline{F}^*_{n,k,i})|}{\sqrt{n}\, \phi^2\{\Phi^{-1}(\overline{F}^*_{n,k,i})\}} \\
&\qquad\qquad\qquad \times (F^*_{n,k} - F_k)^2(X_{i,k}) |\Phi^{-1}\{F_{k'}(X_{i,k'})\}| \\
\lesssim &\sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{|\Phi^{-1}\{F_k(X_{i,k})\}| + 1}{\sqrt{n}\, \{\overline{F}^*_{n,k,i} \wedge (1 - \overline{F}^*_{n,k,i})\}^2} \\
&\qquad\qquad\qquad \times (F^*_{n,k} - F_k)^2(X_{i,k}) |\Phi^{-1}\{F_{k'}(X_{i,k'})\}| \\
\lesssim &\sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{|\Phi^{-1}\{F_k(X_{i,k})\}| + 1}{\sqrt{n}\, \{F_k \wedge (1 - F_k)\}^2(X_{i,k})} \\
&\qquad\qquad\qquad \times (F^*_{n,k} - F_k)^2(X_{i,k}) |\Phi^{-1}\{F_{k'}(X_{i,k'})\}|,
\end{aligned}$$

on the event $E_{4,k,n}$, and hence, by (F.8),

$$\begin{aligned}
|B_{22,kk'}| \lesssim n^{-3/2} \delta^2_{n,p} &\sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{|\Phi^{-1}\{F_k(X_{i,k})\}| + 1}{\{F_k \wedge (1 - F_k)\}(X_{i,k})} \\
&\qquad \times |\Phi^{-1}\{F_{k'}(X_{i,k'})\}| \equiv C_{22,kk'} \qquad \text{(F.40)}
\end{aligned}$$

on the event $E_{4,k,n}$. It is easy to see that

$$\mathrm{E}(C_{22,kk'}) \leq C n^{-1/2} \ln(p) \ln^2(n). \tag{F.41}$$

Next, after some algebra it can be shown that for any integer $r \geq 2$, the rth absolute moment of the summand in $C_{22,kk'}$ is bounded by a multiple of a constant which is uniform over $r \geq 2$, and $r! 2^{3r} \ln^r \{1/(2a_n)\} a_n^{-r+1} = r! \left\{64 \ln^2(1/2a_n) a_n^{-1}\right\} \left\{8 \ln(1/2a_n) a_n^{-1}\right\}^{r-2}$ uniformly over r and n. Then, by Bernstein's inequality, there exists an event $E_{8,kk',n}$ with probability at least $1 - 1/p^4$ on which

$$|C_{22,kk'} - \mathrm{E}C_{22,kk'}| \leq C n^{-1/2} \ln(p) \ln(n). \tag{F.42}$$

Therefore, from (F.40), (F.41) and (F.42) we conclude that on $E_{4,k,n} \cap E_{8,kk',n}$,

$$|B_{22,kk'}| \leq C n^{-1/2} \ln(p) \ln^2(n). \tag{F.43}$$

We now finally collect the above results. By (F.12) and (F.18), we obtain

$$\Xi_{2,n,kk'} = B_{1,kk'} + B_{21,kk'} + B_{22,kk'}$$

$$= B_{1,kk'} + \overline{B}_{21,kk'} + (B_{21,kk'} - \overline{B}_{21,kk'}) + B_{22,kk'},$$

where $\overline{B}_{21,kk'}$ satisfies (F.37). The terms $B_{1,kk'}$, $(B_{21,kk'} - \overline{B}_{21,kk'})$ and $B_{22,kk'}$ in the expansion of $\Xi_{2,n,kk'}$, and then the terms D_1 and D_2 in the expansion of $\Xi_{2,n,kk'}$ are bounded as in (F.17), (F.23), (F.43), (F.38), (F.39) respectively. Together these bounds yield that, on the intersection of some events $E_{n,k}$ and $E_{n,kk'}$ with probabilities at least $1-C/p^3$ and $1-C/p^4$ respectively, the magnitude of the sum of these remainder terms is bounded by $Cn^{-1/2}\{\ln(p)\ln^2(n) + \ln^{3/2}(p)\}$. Taking the intersection of $E_{n,k}$ and $E_{n,kk'}$ over $k, k' \in [p]$ then yields the conclusion of Proposition F.5. □

By an analogous argument, the term $\Xi_{3,n,kk'}$ (F.11) satisfies

$$\Xi_{3,n,kk'} = -\frac{r_{kk'}}{2}\frac{1}{\sqrt{n}}\sum_{i\in[n]}[\Phi^{-1}\{F_{k'}(X_{i,k'})\}^2 - 1] + R_{2,n,kk'},$$

where the remainder term $R_{2,n,kk'}$ satisfies the same bound as $R_{1,n,kk'}$ in Proposition F.5. Finally, the term $\Xi_{4,n,kk'}$ in (F.11), which is a second-order term, is treated by the following proposition. The conclusion of Theorem F.1 then follows.

Proposition F.7. *For the term $\Xi_{4,n,kk'}$ in (F.11), with probability at least $1 - 1/p^2$,*

$$\max_{k,k'\in[p]} |\Xi_{4,n,kk'}| \le Cn^{-1/2}\ln(p)\ln(n). \tag{F.44}$$

Proof. Using Hölder's inequality, we bound $\Xi_{4,n,kk'}$ as

$$|\Xi_{4,n,kk'}| \le \Big[\frac{1}{\sqrt{n}}\sum_{i\in[n]}[\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}]^2\Big]^{1/2} \times \Big[\frac{1}{\sqrt{n}}\sum_{i\in[n]}[\Phi^{-1}\{F^*_{n,k'}(X_{i,k'})\} - \Phi^{-1}\{F_{k'}(X_{i,k'})\}]^2\Big]^{1/2}.$$

For the first term on the right-hand side above, as in (F.12), we again decompose the sum over $i \in [n]$ into the cases $F_k(X_{i,k}) \in A_{1,n}$ and $F_k(X_{i,k}) \in A_{2,n}$. We then apply the mean value theorem to the summands in the second case. We arrive at

$$\frac{1}{\sqrt{n}}\sum_{i\in[n]}[\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}]^2 = B_{5,k} + B_{6,k},$$

say, where

$$B_{5,k} = \frac{1}{\sqrt{n}}\sum_{F_k(X_{i,k})\in A_{1,n}}[\Phi^{-1}\{F^*_{n,k}(X_{i,k})\} - \Phi^{-1}\{F_k(X_{i,k})\}]^2$$

and

$$B_{6,k} = \frac{1}{\sqrt{n}} \sum_{F_k(X_{i,k}) \in A_{2,n}} \frac{1}{\phi^2\{\Phi^{-1}(\overline{F}^*_{n,k,i})\}} (F^*_{n,k} - F_k)^2(X_{i,k}).$$

In the above, for each i, $\overline{F}^*_{n,k,i}$ is a random number strictly between $F^*_{n,k}(X_{i,k})$ and $F_k(X_{i,k})$. For the term $B_{5,k}$, similar to (F.17), there exists an event $E_{9,k,n}$ with probability at least $1 - 1/(2p^3)$ on which $B_{5,k} \leq Cn^{-1/2} \ln(p) \ln(n)$.

For the term $B_{6,k}$, we can follow the derivation of the bound (F.43) for the term $B_{22,kk'}$ in (F.18) (in fact, the present case is easier because of the lack of the function Φ^{-1} in the numerator). We obtain that there exists an event $E_{10,k,n}$ with probability at least $1 - 1/(2p^3)$ on which $|B_{6,k}| \leq Cn^{-1/2} \ln(p) \ln(n)$.

Then, it is easy to see that the bound (F.44) holds on the intersection $\cap_{k \in [p]}(E_{9,k,n} \cap E_{10,k,n})$ with probability at least $1 - 1/p^2$. This concludes the proof of Proposition F.7. □

F.3. Discussion

To the best of our knowledge, this paper is the first to provide a justification for the use of a matrix $\mathbf{\Sigma}_n$ of normal-score rank correlations or van der Waerden correlations to estimate the copula correlation matrix in high-dimensional Gaussian copula models. The analysis of this coefficient faces the double hurdle of an unbounded score function (namely Φ^{-1}) and the non-independence of the Gaussianized observations in (F.3). Nonetheless, in view of the clear superiority of the estimator $\mathbf{\Sigma}_n$ in the fixed-dimensional setting, perhaps it is a bit surprising that it did not receive more attention in high dimensions. Theorem F.1 shows that the estimator $\mathbf{\Sigma}_n$ retains its advantages in the latter regime. Moreover, this result has other immediate and broad practical applications which we briefly describe below.

It should first be observed that in the literature, circumstances often require deviation inequalities on the element-wise maximum norm of the deviation between the target $\mathbf{\Sigma}$ and an estimate thereof, previously often based on Kendall's tau or Spearman's rho. For this purpose but based on the estimator $\mathbf{\Sigma}_n$, the remainder term in (F.5) converges so fast that it can practically be ignored. Next, the leading term in (F.5), which is linear in the efficient influence function, involves only iid, centered, sub-exponential random variables. Hence establishing such deviation inequalities for this term is trivial. Thus Theorem F.1 implies that in all such circumstances, the van der Waerden correlation matrix $\mathbf{\Sigma}_n$ can be used instead. In fact, as mentioned earlier, efficiency gains can be expected from the use of $\mathbf{\Sigma}_n$ due to its established linearity in the efficient influence function. Relevant examples include, but are not limited to, [11, 24, 42] and, in the present context but restricted to the Gaussian copula regression model, Propositions 3.1, 4.2 and Theorem 5.2, for which the estimator $\widehat{\mathbf{\Sigma}}$ in (2.2) of the copula correlation matrix $\mathbf{\Sigma}$ in (2.1) can simply be replaced by $\mathbf{\Sigma}_n$.

Second, it sometimes happens that matrix deviation inequalities stronger than those in terms of the element-wise maximum norm are called for. For instance, in this paper Proposition 3.3 and later in the case of row-sparsity, Theorems 5.4 and 5.7 require a deviation inequality for either the operator norm or the $\|\cdot\|_{\ell_\infty,\ell_2}$ norm. For the leading term in (F.5), this should again be straightforward. In contrast, simply aggregating over the remainder term $R_{n,kk'}$ element-wise may yield suboptimal (though useful) results. Properly treating the remainder term in these cases may require more refined arguments, which we do not pursue further herein.

For completeness, we should mention here that a Winsorized version $\widetilde{\mathbf{\Sigma}}_n$ of the van der Waerden correlation matrix $\mathbf{\Sigma}_n$ was considered by Liu et al. [25], who were able to establish a convergence rate of $n^{-1/4}$, up to logarithm factors. Such a rate is clearly too slow to be competitive. These authors later claimed in [24] that the "efficiency result (from [19]) cannot be generalized to the high-dimensional setting." To establish a satisfactory convergence rate, estimation of the copula correlation matrix in Gaussian copulas then quickly and overwhelmingly switched to the Kendall's tau or the Spearman's rho estimator, whose explicit U-statistic structure with bounded score function makes these estimators more amenable to analysis; see, e.g., [1, 24, 42]. Despite the claim from [25] that the lack of Winsorization "does not lead to accurate inference," Theorem F.1 is in fact obtained without truncation. Nevertheless, it may be that some truncation, perhaps not as heavy as that in [25], may still improve performance.

In [24], a simulation study comparing $\mathbf{\Sigma}_n$ or its Winsorized version $\widetilde{\mathbf{\Sigma}}_n$ to the estimators of $\mathbf{\Sigma}$ based on Kendall's tau and Spearman's rho estimator was carried out. The authors concluded that without contamination, all these estimators performed similarly while with contamination, the estimators based on Kendall's tau and Spearman's rho outperformed $\mathbf{\Sigma}_n$ and $\widetilde{\mathbf{\Sigma}}_n$. We carried out a small simulation study to verify these claims. Specifically we repeated Models 1 and 2 in Section 4.1 in [42], with $n = p = 100$. Without contamination, on average $\mathbf{\Sigma}_n$ outperforms the Kendall's tau and Spearman's rho estimators by 5% in terms of Frobenius norm. With contamination present, the message is mixed: with light, deterministic contamination, the estimator $\mathbf{\Sigma}_n$ can outperform the Kendall's tau and the Spearman's rho estimators by as much as 20%, although in other scenarios the estimator $\mathbf{\Sigma}_n$ may perform less well. While 5% gain may seem modest, in fixed dimensions and in a constrained parametrization (by a lower-dimensional copula parameter $\boldsymbol{\theta}$) such as the Toeplitz matrix model, the efficiency gain for $\boldsymbol{\theta}$ through the one-step estimator that involves $\mathbf{\Sigma}_n$ can sometimes be as large as 400%; see, e.g., Example 5.2 in [37]. We defer the high-dimensional analogy of this phenomenon to future studies.

References

[1] Barber, R. F., and Kolar, M. ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *Ann. Statist. 46* (2018), 3422–3450. MR3852657

[2] Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. MR3185193

[3] Cai, T. T., Liu, W., and Luo, X. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc. 106* (2011), 594–607. MR2847973

[4] Cai, T. T., and Zhang, L. High-dimensional Gaussian copula regression: Adaptive estimation and statistical inference. *Stat. Sinica 28* (2018), 963–993. MR3791096

[5] Cambanis, S., Huang, S., and Simons, G. On the theory of elliptically contoured distributions. *J. Multivariate Anal. 11* (1981), 368–385. MR0629795

[6] Datta, A., and Zou, H. CoCoLasso for high-dimensional error-in-variables regression. *Ann. Statist. 45* (2017), 2400–2426. MR3737896

[7] de la Peña, V., and Giné, E. *Decoupling: From Dependence to Independence. Randomly Stopped Processes. U-statistics and Processes. Martingales and Beyond.* Springer-Verlag, New York, 1999. MR1666908

[8] Devlin, S. J., Gnanadseikan, R., and Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. *Biometrika 62* (1975), 531–545.

[9] Donoho, D. L., and Huo, X. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory 47* (2001), 2845–2862. MR1872845

[10] El Maache, H., and Lepage, Y. Spearman's rho and Kendall's tau for multivariate data sets. In *Mathematical statistics and applications: Festschrift for Constance van Eeden*, vol. 42 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, 2003, pp. 113–130. MR2138289

[11] Fan, J., Xue, L., and Zou, H. Multitask quantile regression under the transnormal model. *J. Amer. Statist. Assoc. 111* (2016), 1726–1735. MR3601731

[12] Fang, K., Kotz, S., and Ng, K. *Symmetric Multivariate and Related Distributions*. Chapman & Hall, London, 1990. MR1071174

[13] Genest, C., Favre, A.-C., Béliveau, J., and Jacques, C. Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research 43* (2007), https://doi.org/10.1029/2006WR005275.

[14] Giné, E., Latała, R., and Zinn, J. Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II (Seattle, WA, 1999)*, vol. 47. Birkhäuser Boston, Boston, MA, 2000, pp. 13–38. MR1857312

[15] Hájek, J., and Šidák, Z. *Theory of Rank Tests.* Prague: Academia, 1967. MR0229351

[16] Huang, J., and Zhang, T. The benefit of group sparsity. *Ann. Statist. 38* (2010), 1978–2004. MR2676881

[17] Hult, H., and Lindskog, F. Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. Appl. Probab. 34* (2002), 587–608. MR1929599

[18] Kendall, M. G., and Gibbons, J. D. *Rank Correlation Methods*, 5th ed. London: Edward Arnold, 1990. MR1079065

[19] Klaassen, C. A. J., and Wellner, J. A. Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli 3* (1997), 55–77. MR1466545

[20] Kruskal, W. H. Ordinal measures of association. *J. Amer. Statist. Assoc. 53* (1958), 814–861. MR0100941

[21] Li, X., Xu, Y., Zhao, T., and Liu, H. Statistical and computational tradeoffs of regularized Dantzig-type estimators. *e-prints* (2015).

[22] Lin, J., Basu, S., Banerjee, M., and Michailidis, G. Penalized maximum likelihood estimation of multi-layered Gaussian graphical models. *J. Mach. Learn. Res. 17* (2016), 1–51. MR3555037

[23] Lindskog, F., McNeil, A. J., and Schmock, U. Kendall's tau for elliptical distributions. In *Credit Risk: Measurement, Evaluation and Management*, G. Bol, G. N. Akhaeizadeh, S. T. Rachev, T. Ridder, and K.-H. Vollmer, Eds. Physica-Verlag, 2003, pp. 149–156. MR1675308

[24] Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. A. High dimensional semiparametric Gaussian copula graphical models. *Ann. Statist. 40* (2012), 2293–2326. MR3059084

[25] Liu, H., Lafferty, J., and Wasserman, L. A. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res. 10* (2009), 2295–2328. MR2563983

[26] Liu, H., Zhang, J., Jiang, X., and Liu, J. The group Dantzig selector. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010), Y. W. Teh and M. Titterington, Eds., vol. 9 of *Proceedings of Machine Learning Research*, PMLR, pp. 461–468.

[27] Loh, P.-K., and Wainwright, M. J. High-dimension regression with noisy and missing data: Provable guarantees with non-convexity. *Ann. Statist. 40* (2012), 1637–1664. MR3015038

[28] Loh, P.-K., and Wainwright, M. J. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res. 16* (2015), 559–616. MR3335800

[29] Lounici, K., Pontil, M., van de Geer, S., and Tsybakov, A. B. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist. 39* (2011), 2164–2204. MR2893865

[30] Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci. 27* (2012), 538–557. MR3025133

[31] Nelsen, R. B. *An Introduction to Copulas*, 2nd ed. Springer, New York, 2006. MR2197664

[32] Obozinski, G., Wainwright, M. J., and Jordan, M. I. Support union recovery in high-dimensional multivariate regression. *Ann. Statist. 39* (2011), 1–47. MR2797839

[33] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized

log-determinant divergence. *Electron. J. Statist. 5* (2011), 935–980. MR2836766

[34] Rothman, A. J., Bickel, P., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electron. J. Statist. 2* (2008), 494–515. MR2417391

[35] Rothman, A. J., Levina, E., and Zhu, J. Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Stat. 19* (2010), 947–962. MR2791263

[36] Ruymgaart, F. H. Asymptotic normality of nonparametric tests for independence. *Ann. Statist. 2* (1974), 892–910. MR0386140

[37] Segers, J., van den Akker, R., and Werker, B. J. M. Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation. *Ann. Statist. 42* (2014), 1911–1940. MR3262472

[38] Shorack, G. R., and Wellner, J. A. *Empirical Processes with Applications to Statistics*. Wiley, New York, 1986. MR0838963

[39] van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998. MR1652247

[40] Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Application*, Y. Eldar and G. Kutyniok, Eds. Cambridge University Press, 2012, pp. 210–268. MR2963170

[41] Wegkamp, M., and Zhao, Y. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli 25* (2016), 1184–1226. MR3449812

[42] Xue, L., and Zou, H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist. 40* (2012), 2541–2571. MR3097612

[43] Yin, J., and Li, H. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat. 5* (2011), 831–851. MR2907129

[44] Yuan, M., and Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B 68* (2006), 49–67. MR2212574

[45] Yuan, M., and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika 94* (2007), 19–35. MR2367824

[46] Zhang, T., and Zou, H. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika 101* (2014), 103–120. MR3180660