ROCKET: Robust Confidence Intervals via Kendall's Tau for Transelliptical Graphical Models

Rina Foygel Barber[†] and Mladen Kolar[‡]

Department of Statistics[†] and Booth School of Business[‡]

The University of Chicago

First draft: Feb 26 20145
This draft: March 21 2015

Abstract

Understanding complex relationships between random variables is of fundamental importance in high-dimensional statistics, with numerous applications in biological and social sciences. Undirected graphical models are often used to represent dependencies between random variables, where an edge between to random variables is drawn if they are conditionally dependent given all the other measured variables. A large body of literature exists on methods that estimate the structure of an undirected graphical model, however, little is known about the distributional properties of the estimators beyond the Gaussian setting. In this paper, we focus on inference for edge parameters in a high-dimensional transelliptical model, which generalizes Gaussian and nonparanormal graphical models. We propose ROCKET, a novel procedure for estimating parameters in the latent inverse covariance matrix. We establish asymptotic normality of ROCKET in ultra high-dimensional setting under mild assumptions, without relying on oracle model selection results. ROCKET requires the same number of samples that are known to be necessary for obtaining a $\sqrt{n}$ consistent estimator of an element in the precision matrix under a Gaussian model. Hence, it is an optimal estimator under a much larger family of distributions. The result hinges on a tight control of the spectral norm of the non-parametric estimator of the correlation matrix, which is of independent interest. Empirically, ROCKET outperforms the nonparanormal and Gaussian models in terms of achieving accurate inference on simulated data. We also compare the three methods on real data (daily stock returns), and find that the ROCKET estimator is the only method whose behavior across subsamples agrees with the distribution predicted by the theory.

Keywords: Graphical model selection; Transelliptical graphical models; Covariance selection; Uniformly valid inference; Post-model selection inference; Rank-based estimation

1 Introduction

Probabilistic graphical models (Lauritzen, 1996) have been widely used to explore complex system and aid scientific discovery in areas ranging from biology and neuroscience to financial modeling and social media analysis. An undirected graphical model consists of a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the set of vertices and E is the set of edges, and a p-dimensional random vector $X = (X_1, \ldots, X_p)^T$ that is Markov with respect to G. In particular, we have that X_a and X_b are conditionally independent given the remaining variables $\{X_c \mid c \in \{1, \ldots, p\} \backslash \{a, b\}\}$ if and only if $\{a, b\} \notin E$. One of the central questions in high-dimensional statistics is estimation of the undirected graph G given n independent realizations of X, as well as quantifying uncertainty of the estimator.

In this paper we focus on (asymptotic) inference for elements in the latent inverse covariance matrix under the semiparametric elliptical copula model (Embrechts et al., 2003; Klüppelberg et al., 2008), also known as the transelliptical model (Liu et al., 2012b). Let $X_1, \ldots, X_n$ be n independent copies of the random vector X that follows a transelliptical distribution,

$$X \sim \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_p), \tag{1.1}$$

where $\Sigma \in \mathbb{R}^p$ is a correlation matrix (that is, $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$), $\xi \in \mathbb{R}$ is a nonnegative random variable with $\mathbb{P}\{\xi = 0\} = 0$, and $f_1, \ldots, f_p$ are univariate, strictly increasing functions. Recall that X follows a transelliptical distribution if the marginal transformation $(f_1(X_1), \ldots, f_p(X_p))$ of X follows a (centered) elliptically contoured distribution with covariance matrix Σ (Fang et al., 1990). Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, also known as the precision matrix; under a Gaussian model, nonzero elements in Ω correspond to pairs of variables that are conditionally dependent, i.e. form an edge in the graph G. Under the model in (1.1), we construct an estimator for a fixed element of the precision matrix, Ω_{ab}, that is asymptotically normal. Furthermore, we construct a confidence interval for the unknown parameter Ω_{ab} that is valid and robust to model selection mistakes. Finally, we construct a uniformly valid hypothesis test for the presence of an edge in the graphical model.

Our main theoretical result establishes that given initial estimates of the regression coefficients for $(f_a(X_a), f_b(X_b)$ on $(f_j(X_j))_{j \neq a,b}$, one can obtain a $\sqrt{n}$-consistent and asymptotically normal estimator for Ω_{ab}. These initial estimators need to converge at a sufficiently fast rate (see Section 3). In particular, we note that we do not require strict sparsity in these regressions, and allow for an error rate that is achievable by known methods such as a nonconvex Lasso (Loh and Wainwright, 2013) (see Section 3.1). To achieve $\sqrt{n}$-consistent rate, our estimator requires the same scaling for sample size n as in the Gaussian case, which is minimax optimal (Ren et al., 2013).

Given accurate initial estimates, in order to construct the asymptotically normal estimator, we prove a key result: that the vector $\mathrm{sign}(X_i - X_{i'})$ is subgaussian at the scale $\mathsf{C}(\Sigma)$ (the condition number of Σ), with dependence on the dimension p coming only through Σ. This result allows us to construct an asymptotically normal estimator by combining the initial regression coefficient estimates with the Kendall's tau rank correlation matrix. In particular, the subgaussianity result allows us to establish a new concentration result on the operator norm of the Kendall's tau correlation matrix that hold with exponentially high probability. This result allows us to uniformly control deviations of quadratic forms involving the Kendall's tau correlation matrix over approximately sparse vectors. These results are of independent interest and could be used to improve recent results of Mitra and Zhang (2014), Wegkamp and Zhao (2013) and Han and Liu (2013). Furthermore, subgaussianity of $\text{sign}(X_i - X_{i'})$ allows us to study properties of penalized rank regression in high-dimensions.

We base our confidence intervals and hypothesis tests on the asymptotically normal estimator of the element Ω_{ab} (see Section 2). We point out that our results hold under milder conditions than those required in Ren et al. (2013), which treats the special case of Gaussian graphical models. Most notably, we give a $\sqrt{n}$-consistent estimator for elements in the precision matrix without requiring strong parametric assumptions.

1.1 Relationship To Literature

Our work contributes to several areas. First, we contribute to the growing literature on graphical model selection in high dimensions. There is extensive literature on the Gaussian graphical model, where it is assumed that $X \sim N(0, \Sigma)$, in which case the edge set E of the graph G is encoded by the non-zero elements of the precision matrix Ω (Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Rothman et al., 2008; Friedman et al., 2008; d'Aspremont et al., 2008; Fan et al., 2009; Lam and Fan, 2009; Yuan, 2010; Cai et al., 2011; Liu and Wang, 2012; Zhao and Liu, 2014). Learning structure of the Ising model based on the penalized pseudo-likelihood was studied in Höfling and Tibshirani (2009), Ravikumar et al. (2010) and Xue et al. (2012). More recently, Yang et al. (2013) studied estimation of graphical models under the assumption that each node's conditional distribution belongs to an exponential family distribution. See also Guo et al. (2011a), Guo et al. (2011b), Lee and Hastie (2012), Cheng et al. (2013), Yang et al. (2012) and Yang et al. (2014) who studied mixed graphical models, where node's conditional distributions are not necessarily all from the same family (for instance, there may be continuous-valued nodes as well as discrete-valued nodes). The parametric Gaussian assumption was relaxed in Liu et al. (2009), where graphic estimation was studied under a Gaussian copula model. More recently, Liu et al. (2012a), Xue and Zou (2012) and Liu et al. (2012b) show that the graph can be recovered in the Gaussian and elliptical semiparametric model class under the same conditions on the sample size n, number of nodes p and the maximum node degree in the graph s as if the estimation was done under the Gaussian assumption. In our paper, we construct a novel $\sqrt{n}$-consistent estimator of an element in the precision matrix without requiring oracle model selection properties.

Second, we contribute to the literature on high-dimensional inference. Recently, there has been much interest on performing valid statistical inference in the high-dimensional setting. Zhang and Zhang (2013), Belloni et al. (2013a), Belloni et al. (2013d), van de Geer et al. (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2013), and Farrell (2013) developed methods for construction of confidence intervals for low dimensional parameters in high-dimensional linear and generalized linear models, as well as hypothesis tests. These methods construct honest, uniformly valid confidence intervals and hypothesis test based on the ℓ_1-penalized estimator in the first stage. Similar results were obtained in the context of the ℓ_1-penalized least absolute deviation and quantile regression (Belloni et al., 2013c,b). Lockhart et al. (2014) study significance of the input variables that enter the model along the lasso path. Lee et al. (2013) and Taylor et al. (2014) perform post-selection inference conditional on the selected model. Liu (2013), Ren et al. (2013) and Chen et al. (2013) construct $\sqrt{n}$-consistent estimators for elements of the precision matrix Ω under a Gaussian assumption. We extend these result to perform valid inference under semiparametric ellitical copula models. In a recent independent work, Gu et al. (2015) propose a procedure for inference under a nonparanormal model. We will provide a detailed comparison in Section 3.

1.2 Notation

Let $[n]$ denote the set $\{1, \ldots, n\}$ and let $\mathbb{I}\{\cdot\}$ denote the indicator function. For a vector $a \in \mathbb{R}^d$, we let $\text{supp}(a) = \{j \ : \ a_j \neq 0\}$ be the support set, and let $||a||_q$, for $q \in [1, \infty)$, be the ℓ_q-norm defined as $||a||_q = (\sum_{i\in[n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $||a||_0 = |\text{supp}(a)|$ and $||a||_\infty = \max_{i\in[n]} |a_i|$.

For a matrix $A \in \mathbb{R}^{n_1\times n_2}$, for sets $S \subset [n_1]$ and $T \subset [n_2]$, we write A_{ST} to denote the $|S| \times |T|$ submatrix of A obtained by extracting the appropriate rows and columns. The sets S and/or T can be replaced by single indices, for example, for $S \subset [n_1]$ and $j \in [n_2]$, A_{Sj} is a $|S|$-length vector. If $A \in \mathbb{R}^{n\times n}$ is a square matrix, for any $T \subset [n]$ we may write A_T to denote the square submatrix A_{TT}.

For a matrix $A \in \mathbb{R}^{n_1\times n_2}$, we use the notation $\mathsf{vec}(A)$ to denote the vector in $\mathbb{R}^{n_1 n_2}$ formed by stacking the columns of A. We denote the Frobenius norm of A by $||A||_{\mathsf{F}}^2 = \sum_{i\in[n_1],j\in[n_2]} A_{ij}^2$, and the operator norm (spectral norm) by $||A||_{\mathsf{op}}$, that is, the largest singular value of A. The norms $||A||_1$ and $||A||_\infty$ are applied entrywise, with $||A||_1 = \sum_{ij} |A_{ij}|$ and $||A||_\infty = \max_{ij} |A_{ij}|$. We write $\mathsf{C}(A)$ to denote the condition number of A, that is, the ratio between the largest and smallest singular values. For two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{r\times s}$, $A\otimes B \in \mathbb{R}^{nr\times ms}$ denotes the Kronecker product, with $(A\otimes B)_{ik,jl} = A_{ij}B_{kl}$. For two matrices of the same size, $A, B \in \mathbb{R}^{n\times m}$, $A\circ B \in \mathbb{R}^{n\times m}$ denotes the Hadamard product (that is, the entrywise product), with $(A\circ B)_{ij} = A_{ij}B_{ij}$. Kronecker products and Hadamard products are defined also for vectors, by treating a vector as a matrix with one column.

Throughout, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, that is, $\Phi(t) = \mathbb{P}\{N(0, 1) \leqslant t\}$.

1.3 Organization of the paper

In Section 2 we introduce Gaussian graphical models and their nonparametric extensions: nonparanormal and transelliptical graphical models. It is illustrated that transelliptical graphical models are useful for modeling dependent tail events, which cannot be modeled with Gaussian and nonparanormal graphical models. We further introduce our asymptotically normal estimator, ROCKET, for edge parameters in a transelliptical model. Our main theoretical result, which establishes distributional properties of the ROCKET estimator, is given in Section 3 together with technical assumptions. Section 3.1 discusses choices of initial estimators. It is shown that the non-convex Lasso estimator can be used under the same conditions used to study the Gaussian case. Section 4 provides an outline of the proof for the main result and the key technical result. Section 5 provides illustrative simulations. An application to S&P 500 stock price closing data is given in Section 6. We conclude the paper with a discussion. Technical proofs are relegated to Appendix.

2 Preliminaries and method

Before introducing our method, we begin with some preliminary definitions and properties of the transelliptical distribution, and related models.

Gaussian and nonparanormal graphical models Suppose that $X = (X_1, \dots, X_p)$ follows a multivariate normal distribution,

$$X \sim N(\mu, \Sigma) \ .$$

A Gaussian graphical model represents the structure of the covariance matrix Σ with a graph, where an edge between nodes a and b indicates that $\Omega_{ab} \neq 0$, where $\Omega = \Sigma^{-1}$ is the precision (inverse covariance) matrix. This model can be generalized by allowing for arbitrary marginal transformations on the variables $X_1, \dots, X_p$. Liu et al. (2009) study the resulting distribution, the nonparanormal model (also known as a Gaussian copula), where we write

$$X \sim \mathsf{NPN}(\Sigma; f_1, \dots, f_p),$$

if the marginally transformed vector $(f_1(X_1), \dots, f_p(X_p))$ follows a (centered) multivariate normal distribution,

$$(f_1(X_1), \dots, f_p(X_p)) \sim N(0, \Sigma) \ .$$

The sparse structure of the underlying graphical model, representing the sparsity pattern in $\Omega = \Sigma^{-1}$, can then be recovered using similar methods as in the Gaussian case. Note that the Gaussian model is a special case of the nonparanormal model (by setting $f_1, \dots, f_p$ each to be the identity function).

Elliptical and transelliptical graphical models The elliptical model is a generalization of the Gaussian graphical model that allows for heavier-tailed dependence between variables. The random vector $X = (X_1, \dots, X_p)$ follows an elliptical distribution with the mean vector $\mu \in \mathbb{R}^p$, covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, and a random variable (the "radius") $\xi \geqslant 0$, denoted by

$$X \sim \mathsf{E}(\mu, \Sigma, \xi) \ ,$$

if we can write $X = \mu + \xi \cdot A \cdot U$, where $AA^\top = \Sigma$ is a Cholesky decomposition of Σ, and where $U \in \mathbb{R}^p$ is a unit vector drawn uniformly at random (independently from the radius ξ). Note that the level sets of this distribution are given by ellipses, centered at μ and with shape determined by Σ. The Gaussian model is a special case of the elliptical model (by taking $\xi \sim \chi_p$).

The transelliptical model (also known as an elliptical copula) combines the elliptical distribution with marginal transformations, much as the nonparanormal distribution applies marginal transformations to a multivariate Gaussian. For a random vector $X \in \mathbb{R}^p$ we write

$$X \sim \mathsf{TE}(\Sigma, \xi; f_1, \dots, f_p)$$

to denote that the marginally transformed vector $(f_1(X_1), \dots, f_p(X_p))$ follows a centered elliptical distribution, specifically,

$$(f_1(X_1), \dots, f_p(X_p)) \sim \mathsf{E}(0, \Sigma, \xi) \ .$$

Here the marginal transformation functions $f_1, \dots, f_p$ are assumed to be strictly increasing. Note that the Gaussian, nonparanormal, and elliptical models are each special cases of this model.

Pearson's rho and Kendall's tau From this point on, we assume for each distribution that $\mu = 0$ and that Σ is a correlation matrix (that is, diagonal elements are equal to one, $\Sigma_{aa} = 1$). In the case of the Gaussian distribution $X \sim N(0, \Sigma)$, the entries of Σ are the (population-level) Pearson's correlation coefficients for each pair of variables, which in this case we can also write as $\Sigma_{ab} = \mathbb{E}[X_a X_b]$. In this setting, we can estimate Σ with the sample covariance.

In the nonparanormal setting, $X \sim \mathsf{NPN}(\Sigma; f_1, \ldots, f_p)$, it is no longer the case that Σ_{ab} is equal to the (population-level) correlation $\mathrm{Corr}(X_a, X_b)$, due to the marginal transformations. However, we can estimate $f_1, \ldots, f_p$ by performing marginal empirical transformations of each X_a to the standard normal distribution. After taking these empirical transformations, Σ can again be estimated via the empirical covariances. Similarly, for the elliptical model $X \sim \mathsf{E}(0, \Sigma, \xi)$, after rescaling so that $\mathbb{E}[\xi^2] = p$ we also have $\Sigma_{ab} = \mathbb{E}[X_a X_b]$. We can therefore again estimate Σ via the empirical covariance.

For the transelliptical distribution, in contrast, this is no longer possible. Taking scaling $\mathbb{E}[\xi^2] = p$ for simplicity, we generalize the calculations above to have

$$\Sigma_{ab} = \mathbb{E}[f_a(X_a) f_b(X_b)] \; .$$

Therefore, if we can estimate the marginal transformations $f_1, \ldots, f_p$, then we can estimate Σ using the empirical covariance of the transformed data. However, unlike the nonparanormal model, estimating $f_1, \ldots, f_p$ is not straightforward. The reason is that, for the elliptical distribution $\mathsf{E}(0, \Sigma, \xi)$, the marginal distributions are not known unless the distribution of the radius ξ is known. Therefore, marginally for each X_a, we cannot estimate f_a because we do not know what should be the marginal distribution after transformation, that is, what should be the marginal distribution of $f_a(X_a)$. (In contrast, in the nonparanormal model, we know that $f_a(X_a)$ is marginally normal.)

As an alternative, Liu et al. (2012b) use the Kendall rank correlation coefficient (Kendall's tau). At the population level, Kendall's tau is given by

$$\tau_{ab} := \tau(X_a, X_b) = \mathbb{E}\left[\mathrm{sign}(X_a - X'_a) \cdot \mathrm{sign}(X_b - X'_b)\right] \; ,$$

where X' is an i.i.d. copy of X. Unlike Pearson's rho, the Kendall's tau coefficient is invariant to marginal transformations: since f_a, f_b are strictly increasing functions, we see that

$$\mathrm{sign}(f_a(X_a) - f_a(X'_a)) \cdot \mathrm{sign}(f_b(X_b) - f_b(X'_b)) = \mathrm{sign}(X_a - X'_a) \cdot \mathrm{sign}(X_b - X'_b) \; .$$

At the sample level, Kendall's tau can be estimated by taking a U-statistic comparing each pair of distinct observations:

$$\widehat{\tau}_{ab} = \frac{1}{\binom{n}{2}} \sum_{1 \leqslant i < i' \leqslant n} \mathrm{sign}(X_{ia} - X_{i'a}) \cdot \mathrm{sign}(X_{ib} - X_{i'b}) \; . \tag{2.1}$$

When X follows an elliptical distribution, Therorem 2 of Lindskog et al. (2003) gives us the following relationship between Kendall's tau and the Pearson's rho coefficients given by the covariance matrix Σ:

$$\Sigma_{ab} = \sin\left(\frac{\pi}{2} \tau_{ab}\right) \text{ for each } a, b \in [p] \; .$$

Since Kendall's tau is invariant to marginal transformations, this identity holds for the transelliptical family as well. For this reason, Liu et al. (2012b) estimate the covariance matrix Σ by

$$\widehat{\Sigma}_{ab} = \sin\left(\frac{\pi}{2}\widehat{\tau}_{ab}\right) \ . \tag{2.2}$$

Note, however, that $\widehat{\Sigma}$ is not necessarily positive semidefinite.

For the remainder of this paper, $\widehat{\Sigma}$ denotes the estimate given here in (2.2). The matrix of the Kendall's tau coefficients is denoted as T, that is, $T_{ab} := \tau_{ab}$, and its empirical estimate (with entries defined as in (2.1)) is denoted as $\widehat{T}$.

Comparing models: tail dependence It is clear that, compared to a Gaussian graphical model, the nonparanormal model allows for data that may be extremely heavy-tailed (in the marginal distributions). A more subtle consideration is the question of tail dependence between two or more of the variables. In particular, the nonparanormal model does not allow for tail dependence between two variables to be any stronger than in the Gaussian distribution itself. Specifically, consider pairwise α-tail dependence between X_a and X_b, given by

$$\mathsf{Tail}_\alpha(X_a, X_b) := \mathrm{Corr}\left(\mathbb{I}\left\{X_a \geqslant q_\alpha^{X_a}\right\}, \mathbb{I}\left\{X_b \geqslant q_\alpha^{X_b}\right\}\right) \ ,$$

where $q_\alpha^{X_a}$ is the α-quantile of the marginal distribution of X_a, and same for X_b. Taking $\alpha \to 1$, this is a measure of the correlation between the extreme right tail of X_a and the extreme right tail of X_b. (Of course, we can also consider the left tail of the distribution of X_a and/or X_b.)

Note that marginal transformations of each variable do not affect this measure, since the quantiles $q_\alpha^{X_a}, q_\alpha^{X_b}$ take these transformations into account. In particular, the nonparanormal distribution has the same tail correlations $\mathsf{Tail}_\alpha(X_a, X_b)$ as the multivariate Gaussian distribution (with the same Σ). In contrast, an elliptical or transelliptical model can exhibit much higher tail correlations. Since real data often exhibits heavy tail dependence between variables, the flexible transelliptical model may be a better fit in many applications.

We demonstrate this behavior with a simple example in Figure 1. Here we take

$$X = (X_1, X_2) \sim \mathsf{E}(0, \Sigma, \xi) \text{ with } \Sigma = \begin{pmatrix} 1 & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 \end{pmatrix} \ , \tag{2.3}$$

where $\xi \sim \chi_2 \cdot \sqrt{d}/\chi_d$ for $d \in \{0.1, 1, 5, 10, \infty\}$, corresponding to a multivariate t-distribution with d degrees of freedom (note that $d = \infty$ is equivalent to taking $X \sim N(0, \Sigma)$). Note that at $\alpha = 0.5$, the tail correlation $\mathsf{Tail}_\alpha(X_1, X_2)$ is equal to the Kendall's tau coefficient $\tau(X_1, X_2) = \frac{2}{\pi}\arcsin(\Sigma_{12}) = 0.5$. Figure 1 shows that, as $\alpha \to 1$, the tail correlation decreases towards zero for the normal distribution ($d = \infty$) but grows for low values of d.

2.1 ROCKET: an asymptotically normal estimator

Suppose that our data points X_i are drawn i.i.d. from a transelliptical distribution with covariance matrix Σ. We would like to perform inference on a particular entry of the precision matrix $\Omega = \Sigma^{-1}$, specifically, we are interested in producing a confidence interval for Ω_{ab} where $a \neq b \in \{1, \dots, p\}$ is a prespecified node pair.

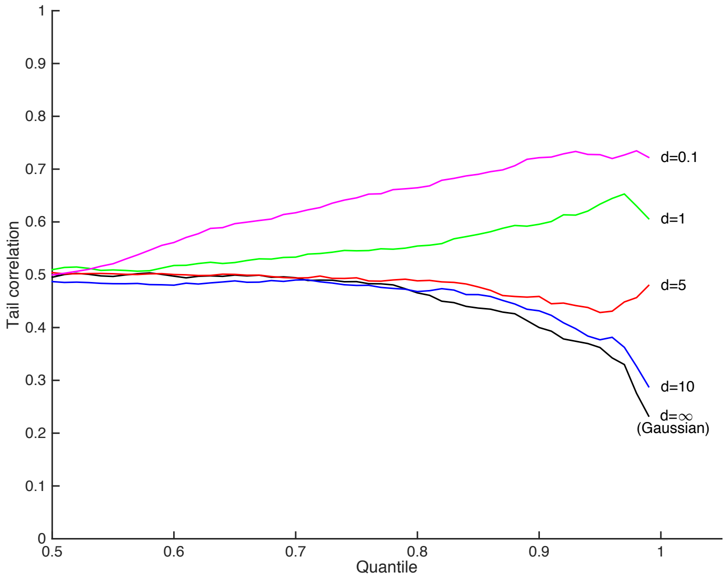

Figure 1: Tail dependence for normal and elliptical distributions on $\mathbb{R}^2$. Data is generated as in (2.3). The figure displays $\mathsf{Tail}_\alpha(X_1, X_2)$, estimated empirically from a sample size $n = 20000$.

To move towards constructing a confidence interval, we introduce a few definitions and calculations. First, let $I = \{1, \ldots, p\}\backslash\{a, b\}$, and observe that by block-wise matrix inversion, we can calculate the $\{a, b\} \times \{a, b\}$ sub-block of Ω as follows:

$$\Omega_{ab,ab} = \left(\Sigma_{ab,ab} - \Sigma_{ab,I}\Sigma_I^{-1}\Sigma_{I,ab}\right)^{-1} . \tag{2.4}$$

Define $\gamma_a = \Sigma_I^{-1}\Sigma_{Ia}$ and $\gamma_b = \Sigma_I^{-1}\Sigma_{Ib}$ (in the Gaussian graphical model setting, these are the regression coefficients when $f_a(X_a)$ or $f_b(X_b)$ is regressed on $\{f_j(X_j) : j \in I\}$). We then have

$$\Sigma_{ab,I}\Sigma_I^{-1}\Sigma_{I,ab} = (\gamma_a \ \gamma_b)^\top\Sigma_{I,ab} = \Sigma_{I,ab}^\top(\gamma_a \ \gamma_b) = (\gamma_a \ \gamma_b)^\top\Sigma_I(\gamma_a \ \gamma_b) .$$

We can therefore rewrite (2.4) as follows (this somewhat redundant formulation will allow for a favorable cancellation of error terms later on):

$$\Theta := (\Omega_{ab,ab})^{-1} = \Sigma_{ab,ab} - (\gamma_a \ \gamma_b)^\top\Sigma_{I,ab} - \Sigma_{I,ab}^\top(\gamma_a \ \gamma_b) + (\gamma_a \ \gamma_b)^\top\Sigma_I(\gamma_a \ \gamma_b) . \tag{2.5}$$

We abuse notation and index the entries of Θ with the indices a and b, that is, we denote Θ as lying in $\mathbb{R}^{\{a,b\}\times\{a,b\}}$ rather than $\mathbb{R}^{2\times 2}$.

Next, we define an oracle estimator of Θ, defined by plugging the *true* values of γ_a and γ_b and the *empirical* estimate of Σ (given in (2.2)) into (2.5) above:

$$\widetilde{\Theta} = \widehat{\Sigma}_{ab,ab} - (\gamma_a \ \gamma_b)^\top\widehat{\Sigma}_{I,ab} - \widehat{\Sigma}_{I,ab}^\top(\gamma_a \ \gamma_b) + (\gamma_a \ \gamma_b)^\top\widehat{\Sigma}_I(\gamma_a \ \gamma_b) . \tag{2.6}$$

Later on (in Theorem 4.1), we will show that due to standard results on the theory of U-statistics, this oracle estimator is asymptotically normal. If $\widetilde{\Theta}$ were known, then, we would have achieved our goal for inference in this model, as $\widetilde{\Omega}_{ab} = \left(\widetilde{\Theta}^{-1}\right)_{ab}$ weakly converges to a Normal random variable centered at Ω_{ab} with variance S_{ab}/n.

Of course, in practice we do not know the true values of γ_a and γ_b, and must instead use some available estimators, denoted by $\breve{\gamma}_a$ and $\breve{\gamma}_b$ (we discuss how to obtain these preliminary estimates later on). Given the estimators of the regression vectors, we then define our estimator of Θ as follows:

$$\breve{\Theta} = \widehat{\Sigma}_{ab,ab} - (\breve{\gamma}_a\ \breve{\gamma}_b)^\top \widehat{\Sigma}_{I,ab} - \widehat{\Sigma}_{I,ab}^\top (\breve{\gamma}_a\ \breve{\gamma}_b) + (\breve{\gamma}_a\ \breve{\gamma}_b)^\top \widehat{\Sigma}_I (\breve{\gamma}_a\ \breve{\gamma}_b)\ . \tag{2.7}$$

Since we are interested in Ω_{ab} rather than in the matrix Θ, as a final step we define our estimator

$$\breve{\Omega}_{ab} = \left(\breve{\Theta}^{-1}\right)_{ab}\ . \tag{2.8}$$

In order to make inference about Ω_{ab}, we approximate the distribution of $\breve{\Omega}_{ab}$. Let

$$\breve{\mathsf{Err}}_{ab} = \frac{\breve{\Omega}_{ab} - \Omega_{ab}}{\breve{S}_{ab}}\ , \tag{2.9}$$

be the studentized error, where the normalization term $\breve{S}_{ab}$ is defined below. First, define the (random) kernel

$$\breve{g}(X, X') = \operatorname{sign}(X - X')^\top \left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}\widehat{T}\right)\right) \operatorname{sign}(X - X')\ ,$$

where

$$\breve{u}_a = 1, \breve{u}_b = 0, \breve{u}_I = -\breve{\gamma}_a \text{ and } \breve{v}_a = 0, \breve{v}_b = 1, \breve{v}_I = -\breve{\gamma}_b\ .$$

(Note that we have defined $\breve{u}$ and $\breve{v}$ so that $\breve{\Theta}_{ab} = \breve{u}^\top \widehat{\Sigma} \breve{v}$.) Then define

$$\breve{S}_{ab} = \frac{\pi}{\det(\breve{\Theta})} \cdot \sqrt{\frac{1}{n}\sum_i \left(\frac{1}{n-1}\sum_{i' \neq i} \breve{g}(X_i, X_{i'}) - \mathsf{mean}(\breve{g})\right)^2} \text{ where } \mathsf{mean}(\breve{g}) = \frac{1}{\binom{n}{2}} \sum_{i < i'} \breve{g}(X_i, X_{i'})\ .$$

We will see later on that $\breve{S}_{ab}^2$ estimates the variance of $\breve{\Theta}_{ab}$ and that the expression above arises naturally from the theory of U-statistics.

Our main result, Theorem 3.5 below, will prove that $\sqrt{n} \cdot \breve{\mathsf{Err}}_{ab}$ follows a distribution that is approximately standard normal. Therefore, an approximate $(1-\alpha)$-confidence interval for Ω_{ab} is given by

$$\breve{\Omega}_{ab} \pm z_{\alpha/2} \cdot \frac{\breve{S}_{ab}}{\sqrt{n}}\ , \tag{2.10}$$

where $z_{\alpha/2}$ is the appropriate quantile of the normal distribution, that is, $\mathbb{P}\left\{N(0,1) > z_{\alpha/2}\right\} = \alpha/2$. In order to establish the asymptotic normality of $\sqrt{n} \cdot \breve{\mathsf{Err}}_{ab}$, we will show that $\sqrt{n}||\breve{\Theta} - \widetilde{\Theta}||_\infty = o_P(1)$ and that $\left|\breve{S}_{ab} S_{ab}^{-1} - 1\right| = o_P(1)$.

Notation for fixed vs random quantities From this point on, as much as possible throughout the main body of the paper, quantities that depend on the data and depend on the initial estimates $\check{\gamma}_a, \check{\gamma}_b$ are denoted with a "check" accent, for example, $\check{\Theta}$. Quantities that depend on the data, but do not depend on $\check{\gamma}_a, \check{\gamma}_b$, are denoted with a "hat" accent, for example, $\widehat{\Sigma}$. Any quantities with neither a "hat" nor a "check" are population quantities, that is, they are not random. Two important exceptions are the data itself, $X_1, \dots, X_n$, and the oracle estimator, $\widetilde{\Theta}$, which is of course data-dependent (but does not depend on $\check{\gamma}_a, \check{\gamma}_b$).

3 Main results

In this section, we give a theoretical result showing that the confidence interval constructed in (2.10) has asymptotically the correct coverage probability, as long as we have reasonably accurate estimators of $\gamma_a = \Sigma_I^{-1}\Sigma_{Ia}$ and $\gamma_b = \Sigma_I^{-1}\Sigma_{Ib}$. Our asymptotic result considers a problem whose dimension $p_n \geqslant 2$ grows with the sample size n. We also allow for the sparsity level in the true inverse covariance matrix $\Omega \in \mathbb{R}^{p_n \times p_n}$ to grow.[1] We use k_n to denote an approximate bound on the sparsity in each column of Ω (details given below).

We begin by stating several assumptions on the distribution of the data and on the initial estimators $\check{\gamma}_a$ and $\check{\gamma}_b$. All of the constants appearing in these assumptions should be interpreted as values that do not depend on the dimensions (n, p_n, k_n) of the problem.

Assumption 3.1. *The data points $X_1, \dots, X_n \in \mathbb{R}^{p_n}$ are i.i.d. draws from a transelliptical distribution,*

$$X_i \overset{iid}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \dots, f_{p_n}) ,$$

where $f_1, \dots, f_{p_n}$ are any monotone functions, $\xi \geqslant 0$ is any random variable with $\mathbb{P}\{\xi = 0\} = 0$, and the covariance matrix $\Sigma \in \mathbb{R}^{p_n \times p_n}$ is positive definite, with $\mathrm{diag}(\Sigma) = \mathbf{1}$ and bounded condition number,

$$\mathsf{C}(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leqslant C_{\mathsf{cov}} ,$$

for some constant C_{cov}.

Assumption 3.2. *The a-th and b-th columns of the true inverse covariance Ω, denoted by Ω_a and Ω_b, are approximately k_n-sparse, that is, they satisfy*

$$||\Omega_a||_1 \vee ||\Omega_b||_1 \leqslant C_{\mathsf{sparse}}\sqrt{k_n} ,$$

for some constant C_{sparse}.

Assumption 3.3. *For some constant C_{est} and for some $\delta_n > 0$, with probability at least $1 - \delta_n$, the preliminary estimates $\check{\gamma}_a$ and $\check{\gamma}_b$ of the vectors γ_a and γ_b satisfy*

$$||\check{\gamma}_a - \gamma_a||_2 \vee ||\check{\gamma}_b - \gamma_b||_2 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n \log(p_n)}{n}} \text{ and } ||\check{\gamma}_a - \gamma_a||_1 \vee ||\check{\gamma}_b - \gamma_b||_1 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}} . \quad (3.1)$$

[1] While Σ, Ω, etc, all depend on the sample size n since the dimension of the problem grows, we abuse notation and do not write Σ_n, Ω_n, etc; the dependence on n is implicit.

Assumption 3.4. *Define the kernel*

$$h(X, X') = \text{sign}(X - X') \otimes \text{sign}(X - X') \in \mathbb{R}^{p_n^2}$$

and let

$$h_1(X) = \mathbb{E}\left[h(X, X') \mid X\right] .$$

Define $\Sigma_h = \mathsf{Var}(h(X, X'))$ *and* $\Sigma_{h_1} = \mathsf{Var}(h_1(X))$, *where* $X, X' \overset{iid}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_{p_n})$. *Then for some constant* $C_{\mathsf{kernel}} > 0$,[2]

$$C_{\mathsf{kernel}} \cdot \Sigma_h \preceq \Sigma_{h_1} \preceq \Sigma_h .$$

Assumption 3.1 assumes that the smallest and largest eigenvalues of the correlation matrix Σ are bounded away from zero and infinity, respectively. This assumption is commonly assumed in the literature on learning structure of probabilistic graphical models (Ravikumar et al., 2011; Liu et al., 2009, 2012a). Assumption 3.2 does not require that the precision matrix Ω be exactly sparse, which is commonly assumed in the literature on exact graph recovery (see, for example, Ravikumar et al., 2011), but only requires that rows Ω_a and Ω_b have the ℓ_1 norm that does not grow too fast. Note that if Ω_c, for $c = a, b$, is k_n-sparse vector, then

$$||\Omega_c||_1 \leqslant \sqrt{k_n}||\Omega_c||_2 \leqslant \sqrt{k_n}\lambda_{\max}(\Omega) \leqslant C_{\mathsf{cov}}\sqrt{k_n}$$

and we could then set $C_{\mathsf{sparse}} = C_{\mathsf{cov}}$. Assumption 3.3 is a high-level condition, which assumes existence of initial estimators of γ_a and γ_b that converge at a fast enough rate. In the next section, we will see that Assumption 3.1 together with a stronger version of Assumption 3.2 are sufficient for Assumption 3.3 to be satisfied with a specific estimator that is efficient to compute. Finally, Assumption 3.4 is imposed to allow for estimation of the asymptotic variance $\check{\Omega}_{ab}$.

We now state our main result.

Theorem 3.5. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, there exists a constant* C_{converge}, *depending on* $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{est}}, C_{\mathsf{kernel}}$ *but not on the dimensions* (n, p_n, k_n) *of the problem, such that*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \sqrt{n} \cdot \frac{\check{\Omega}_{ab} - \Omega_{ab}}{\check{S}_{ab}} \leqslant t \right\} - \Phi(t) \right| \leqslant C_{\mathsf{converge}} \cdot \sqrt{\frac{k_n^2 \log^2(p_n)}{n}} + \frac{1}{p_n} + \delta_n .$$

We note that the result holds uniformly over a large class of data generating processes satisfying Assumptions 3.1,3.2, 3.3, and 3.4, which are rather weak assumptions. We emphasize that the result holds without requiring exact model selection or oracle properties, which hold only for restrictive sequences of data generating processes. For example, we do not require the "beta-min" condition (that is, a lower bound on $|\Omega_{ab}|$ for all true edges) or any incoherence conditions (Bühlmann and van de Geer, 2011), which may be implausible in practice. Instead of requiring perfect model selection, we only require estimation consistency as given in Assumption 3.3.

As an immediate corollary, we see that the confidence interval constructed in (2.10) is asymptotically correct:

[2] Here we use the positive semidefinite ordering on matrices, that is, $A \succeq B$ if $A - B \succeq 0$. Note that the second part of the inequality, $\Sigma_{h_1} \preceq \Sigma_h$, is always true by the law of total variance.

Corollary 3.6. *Under the assumptions and notation of Theorem 3.5, the* $(1-\alpha)$*-confidence interval constructed in* (2.10) *fails to cover the true parameter* Ω_{ab} *with probability no higher than*

$$\alpha + 2\left[C_{\mathsf{converge}} \cdot \sqrt{\frac{k_n^2 \log^2(p_n)}{n}} + \frac{1}{p_n} + \delta_n\right] .$$

Again this result holds uniformly over a large class of data generating distributions.

Theorem 3.5 is striking as it shows that we can form an asymptotically normal estimator of Ω_{ab} under the transelliptical distribution family with the sample complexity $n = \Omega\left(k_n^2 \log^2(p_n)\right)$. This sample size requirement was shown to be optimal for obtaining an asymptotically normal estimator of an element in a precision matrix from multivariate normal data (Ren et al., 2013). More precisely, let

$$\mathcal{G}_0(M, k_n) = \left\{ \begin{array}{c} \Omega = (\Omega_{ab})_{a,b\in[p]} \;:\; \max_{a\in[p]} \sum_{b\neq a} \mathbb{I}\{\Omega_{ab} \neq 0\} \leqslant k_n, \\ \text{and } M^{-1} \leqslant \lambda_{\min}(\Omega) \leqslant \lambda_{\max}(\Omega) \leqslant M. \end{array} \right\}$$

where M is a constant greater than one. Then Theorem 1 in Ren et al. (2013) states that

$$\inf_{a,b} \inf_{\breve{\Omega}_{ab}} \sup_{\mathcal{G}_0(M,k_n)} \mathbb{P}\left\{\left|\breve{\Omega}_{ab} - \Omega_{ab}\right| \geqslant \epsilon_0 \left(n^{-1} k_n \log(p_n) \vee n^{-1/2}\right)\right\} \geqslant \epsilon_0$$

and, therefore, our estimator is rate optimal.

At this point, it is also worth mentioning the result of Gu et al. (2015), who study inference under Gaussian copula graphical models. They base their inference procedure on decorrelating a pseudo score function for the parameter of interest and showing that it is normally distributed. Their main result, stated in Theorem 4.10, requires the sample size to satisfy

$$k_n^3 M^6 \left(\frac{\log(p_n)}{n}\right)^{3/2} + k_n^2 M^3 \frac{\log(p_n)}{n} = o\left(n^{-1/2}\right)$$

where $M = \max_{a\in[p]} \sum_{b\in[p]} |\Omega_{ab}|$. As M can be potentially as large as $\sqrt{k_n}$, it is immediately clear that our result achieves much better scaling on the sample size.

3.1 Initial estimators

The validity of our inference method relies in part on the accuracy of the initial estimators $\breve{\gamma}_a$ and $\breve{\gamma}_b$, which are assumed to satisfy error bounds with high probability as stated in Assumption 3.3—that is, with high probability, we have

$$||\breve{\gamma}_a - \gamma_a||_2 \vee ||\breve{\gamma}_b - \gamma_b||_2 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n \log(p_n)}{n}} \text{ and } ||\breve{\gamma}_a - \gamma_a||_1 \vee ||\breve{\gamma}_b - \gamma_b||_1 \leqslant C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}} ,$$

where C_{est} is some constant. Below, we will prove that these required error rates can be obtained, under an additional sparsity assumption, by the Lasso estimators

$$\breve{\gamma}_c = \mathop{\mathrm{argmin}}_{\gamma\in\mathbb{R}^I;||\gamma||_1\leqslant C_{\mathsf{cov}}\sqrt{2k_n}} \left\{\frac{1}{2}\gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ic} + \lambda||\gamma||_1\right\} \tag{3.2}$$

for each $c = a, b$, when the penalty parameter λ is chosen appropriately. In fact, these optimization problems may not be convex, because $\widehat{\Sigma}_I$ will not necessarily be positive semidefinite. Loh and Wainwright (2013) developed theory for this nonconvex high-dimensional setting, which can be applied to our problem to show that Assumption 3.3 is guaranteed to hold with high probability. In particular, any local minimizers of the two optimization problems will satisfy requirements of Assumption 3.3 and, therefore, we only need to be able to run optimization algorithms that find local minima.

We now turn to proving that any local minima for (3.2) for $c = a, b$ will satisfy the required error rates of Assumption 3.3. To proceed, we will use the theoretical results of Loh and Wainwright (2013), which gives a theory for local minimizers of nonconvex regularized objective functions. We specialize their main result to our setting.

Theorem 3.7 (Adapted from Loh and Wainwright (2013, Theorem 1))**.** *Consider any $n, p \geqslant 1$, any $A \in \mathbb{R}^{p\times p}$ and $z \in \mathbb{R}^p$, and any k-sparse $x^\star \in \mathbb{R}^d$ with $||x^\star||_1 \leqslant R$. Suppose that A satisfies restricted strong convexity conditions*

$$v^\top A v \geqslant \alpha_1 ||v||_2^2 - \tau_1 ||v||_1^2 \cdot \frac{\log(p)}{n} \,. \tag{3.3}$$

If

$$n \geqslant \frac{16R^2\tau_1 \max\{\alpha_1, \tau_1\} \log(p)}{\alpha_1^2} \tag{3.4}$$

and

$$\max\left\{4||Ax^\star - z||_\infty, 4\alpha_1\sqrt{\frac{\log(p)}{n}}\right\} \leqslant \lambda \leqslant \frac{\alpha_1}{6R} \tag{3.5}$$

then for any $\breve{x}$ that is a local minimum of the objective function $\frac{1}{2}x^\top Ax - x^\top z + \lambda||x||_1$ over the set $\{x \in \mathbb{R}^d : ||x||_1 \leqslant R\}$, it holds that

$$||\breve{x} - x^\star||_2 \leqslant \frac{1.5\lambda\sqrt{k}}{\alpha_1} \text{ and } ||\breve{x} - x^\star||_1 \leqslant \frac{6\lambda k}{\alpha_1} \,.$$

We apply Theorem 3.7 to our problem of estimating γ_a and γ_b under a setting where the true regression coefficient vectors γ_a and γ_b are exactly sparse.

Corollary 3.8. *Suppose that Assumption 3.1 holds. Assume additionally that the columns Ω_a, Ω_b of the true inverse covariance $\Omega = \Sigma^{-1}$ are k_n-sparse. Then there exist constants $C_{\mathsf{sample}}, C_{\mathsf{Lasso}}$, depending on C_{cov} but not on (n, k_n, p_n), such that if $n \geqslant C_{\mathsf{sample}} k_n \log(p_n)$ then, with probability at least $1 - \frac{1}{2p_n}$, any local minimizer $\breve{\gamma}_a$ of the objective function*

$$\frac{1}{2}\gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ia} + \lambda||\gamma||_1$$

over the set $\{\gamma \in \mathbb{R}^I : ||\gamma||_1 \leqslant C_{\mathsf{cov}}\sqrt{2k_n}\}$ satisfies

$$||\breve{\gamma}_a - \gamma_a||_2 \leqslant 3\sqrt{2}C_{\mathsf{cov}}\lambda\sqrt{k_n} \text{ and } ||\breve{\gamma}_a - \gamma_a||_1 \leqslant 24C_{\mathsf{cov}}\lambda\sqrt{k_n} \,.$$

where we choose $\lambda = C_{\mathsf{Lasso}} \cdot \sqrt{\frac{\log(p_n)}{n}}$. The same result holds for estimating γ_b.

Using this corollary, we see that a local minimizer of (3.2), $\check{\gamma}_c$, satisfies Assumption 3.3 with $\delta_n = \frac{1}{p_n}$ and $C_{\mathsf{est}} = 24 C_{\mathsf{cov}} C_{\mathsf{Lasso}}$.

To prove that this corollary follows from Loh and Wainwright (2013)'s result (Theorem 3.7), it is sufficient to check that the restricted strong convexity condition (3.3) holds, with high probability for the matrix $\widehat{\Sigma}_I$, and then compute the necessary values for λ and the other parameters of Theorem 3.7. The proof is technical and relies on novel results on concentration of the Kendall's tau correlation matrix. Details are given in Appendix C.

We have provided sufficient condition for a local minimizer of (3.2) to satisfy Assumption 3.3, however, many other estimators can be used as initial estimators. For example, one could use the Dantzig selector (Candés and Tao, 2007). Potential benefits of the Dantzig selector over the optimization program in (3.2) are two fold. First, the optimization program is convex even when $\widehat{\Sigma}_I$ is not positive semi-definite. Second, one does not need to know an upper bound R on the ℓ_1 norm of Ω_c for $c = a, b$. Using the techniques similar to those used to prove Corollary 3.8, we can also prove that Assumption 3.3 holds when the Dantzig selector is used as an initial estimator.

In practice, for ease of computation, we remove the constraint on $||\gamma||_1$ in each optimization problem (3.2) for $c = a, b$. Furthermore, we have found that in simulations, using the Lasso for model selection, and then refitting without a penalty, leads to better empirical performance: specifically, for each $c = a, b$, we first fit

$$\check{\gamma}_c^{\mathsf{Lasso}} = \operatorname*{argmin}_{\gamma \in \mathbb{R}^I} \left\{ \frac{1}{2} \gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ia} + \lambda ||\gamma||_1 \right\}$$

(or, more precisely, find a local minimum of this nonconvex optimization problem). We then extract the combined support of these two solution, $\check{J} = \mathsf{supp}(\check{\gamma}_a^{\mathsf{Lasso}}) \cup \mathsf{supp}(\check{\gamma}_b^{\mathsf{Lasso}})$, and refit the coefficients using least-squares:

$$\check{\gamma}_c = \left(\widehat{\Sigma}_{\check{J}} \right)^{-1} \widehat{\Sigma}_{\check{J}c} \text{ for } c = a, b \ .$$

Adapting the proof of Belloni and Chernozhukov (2013), we can also rigorously prove that the refitted estimators also satisfy the Assumption 3.3.

4 Main technical tools

In this section, we outline the proof of Theorem 3.5 and state the key technical result that establishes that sign-subgaussianity property of X that follows a transelliptical distribution. We also illustrate an application of this technical result to establishing a bound on $\widehat{\Sigma} - \Sigma$.

4.1 Sketch of proof for main result

The proof of Theorem 3.5 follows three steps:

- Step 1: prove that the distribution of $\widetilde{\Theta}_{ab}$, the oracle estimator of Θ_{ab}, is asymptotically normal, with

$$\sqrt{n} \cdot \frac{\widetilde{\Theta}_{ab} - \Theta_{ab}}{S_{ab} \det(\Theta)} \to N(0, 1)$$

where S_{ab} is the asymptotic variance of $\widetilde{\Omega}_{ab}$. (Explicit form of S_{ab} is given in the proof of Theorem 4.1.)

- Step 2: prove that the difference between the estimator and the oracle estimator, $\breve{\Theta} - \widetilde{\Theta}$, converges to zero at a fast rate, and that the variance estimator $\breve{S}_{ab}$ converges to S_{ab} at a fast rate.
- Step 3: combining the two steps above, prove that of $\breve{\Omega}_{ab}$ is an asymptotically normal estimator of Ω_{ab}.

The detailed proofs for each step are found in Appendix B. Here, we outline the main results for each step.

Step 1 establishes the Berry-Esseen type bound for the centered and normalized oracle estimator

$$\sqrt{n}\left(S_{ab}\det(\Theta)\right)^{-1}\left(\widetilde{\Theta}_{ab} - \Theta_{ab}\right).$$

We approximate the oracle estimator $\widetilde{\Theta}_{ab}$ by a linear function of the Kendall's tau statistic $\widehat{T}$, which is a U-statistic of the data. We prove that the variance of the linear approximation is bounded away from zero and apply existing results on convergence of U-statistics. The following result is proved in Appendix B.2.

Theorem 4.1. *Suppose that Assumptions 3.1, 3.2, and 3.4 hold. Then there exist constants $C_{\mathsf{normal}}, C_{\mathsf{variance}}$ depending on $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{kernel}}$ but not on (n, p_n, k_n), such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{n}\cdot\frac{\widetilde{\Theta}_{ab} - \Theta_{ab}}{S_{ab}\cdot\det(\Theta)} \leqslant t\right\} - \Phi(t)\right| \leqslant C_{\mathsf{normal}}\cdot\frac{k_n\log(p_n)}{\sqrt{n}} + \frac{1}{2p_n},$$

where S_{ab} is defined in the proof and satisfies $S_{ab}\cdot\det(\Theta) \geqslant C_{\mathsf{variance}} > 0$.

Step 2 contains the main challenge of this problem, since it requires strong results on the concentration properties of the Kendall's tau estimator $\widehat{\Sigma}$ of the covariance matrix Σ. The main ingredient for this step is a new result on "sign-subgaussianity", that is, proving that the signs vector $\mathrm{sign}(X_i - X_{i'})$ is subgaussian for i.i.d. observations $X_i, X_{i'}$. Our results on sign-subgaussianity are discussed in Section 4.2 and their application to concentration of $\widehat{\Sigma}$ around Σ is given in Section 4.3. Using these tools, we are able to prove the following theorem (proved in Appendix B.3):

Theorem 4.2. *Suppose that Assumptions 3.1, 3.2, and 3.3 hold. Then there exists a constant C_{oracle}, depending on $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{est}}$ but not on (n, p_n, k_n), such that, if*[3]

$$n \geqslant 15k_n\log(p_n)$$

then, with probability at least $1 - \frac{1}{2p_n} - \delta_n$,

$$||\breve{\Theta} - \widetilde{\Theta}||_\infty \leqslant C_{\mathsf{oracle}}\cdot\frac{k_n\log(p_n)}{n}$$

and

$$\left|\breve{S}_{ab}\cdot\det(\breve{\Theta}) - S_{ab}\cdot\det(\Theta)\right| \leqslant C_{\mathsf{oracle}}\cdot\sqrt{\frac{k_n^2\log(p_n)}{n}}.$$

[3] Note that the additional condition $n \geqslant 15k_n\log(p_n)$ can be assumed to hold in our main result Theorem 3.5, since if this inequality does not hold, then the claim in Theorem 3.5 is trivial.

Finally, Step 3 simply involves tracking how the errors in $\check{\Theta}$ and in $\check{S}_{ab}$ affect the final distribution, and proving that these errors have a neglible effect relative to the (approximately) standard normal error of the oracle estimator. Details are given in Appendix B.4.

4.2 Sign-subgaussian random vectors

Recall the definition of a subgaussian random vector:

Definition 4.3. *A random vector $X \in \mathbb{R}^p$ is C-subgaussian if, for any fixed vector $v \in \mathbb{R}^p$, it holds that*

$$\mathbb{E}\left[e^{v^\top X}\right] \leqslant e^{C\cdot||v||_2^2/2} \,.$$

For graphical models where the data points X_i come from a subgaussian distribution, the sample covariance matrix $\frac{1}{n}\sum_i(X_i - \overline{X})(X_i - \overline{X})^\top$, with $\overline{X} = \frac{1}{n}\sum_i X_i$, is known to concentrate near the population covariance, as measured by different norms. For example, elementwise convergence of the sample covariance to the population covariance, that is, convergence in $||\cdot||_\infty$, is sufficient to establish rates of convergence for the graphical Lasso, CLIME or graphical Dantzig selector for estimating the sparse inverse covariance (Ravikumar et al., 2011; Cai et al., 2011; Yuan, 2010). Similar results can be obtained also for the transelliptical family, since $||\widehat{T} - T||_\infty \leqslant C\sqrt{\log(p)/n}$ and hence $||\widehat{\Sigma} - \Sigma||_\infty \leqslant C\sqrt{\log(p)/n}$, as was shown in Liu et al. (2012a) and Liu et al. (2012b). However, in order to construct asymptotically normal estimators for the elements of the precision matrix, stronger results are needed about the convergence of the sample covariance to the population covariance (Ren et al., 2013). In particular, a result on convergence in spectral norm, uniformly over all sparse submatrices, is required. One can relate the convergence in the elementwise ℓ_∞ norm to (sparse(spectral norm convergence, however, this would lead to suboptimal sample size. One way to obtain a tight bound on the (sparse) spectral norm convergence is by utilizing subgaussianity of the data points X_i. This is exactly what we proceed to establish.

Recall from (2.2) the Kendall's tau estimator of the covariance,

$$\widehat{\Sigma} = \sin\left(\frac{\pi}{2}\widehat{T}\right) \text{ where } \widehat{T} = \frac{1}{\binom{n}{2}}\sum_{i<i'} \mathrm{sign}(X_i - X_{i'})\,\mathrm{sign}(X_i - X_{i'})^\top \,.$$

Therefore, it is crucial to determine whether the vector $\mathrm{sign}(X_i - X_{i'})$ is itself subgaussian, with the variance proxy that depends on the ambient dimension p_n only through $C(\Sigma)$.[4] Using past results on elliptical distributions, we can reduce to a simpler case using the arguments of Lindskog et al. (2003) (proved in Appendix D.2):

Lemma 4.4. *Let*

$$X, X' \overset{iid}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \ldots, f_p) \,.$$

Suppose that Σ is positive definite, and that $\xi > 0$ with probability 1. *Then* $\mathrm{sign}(X - X')$ *is equal in distribution to* $\mathrm{sign}(Z)$, *where $Z \sim N(0, \Sigma)$.*

[4] Note that $v^\top \mathrm{sign}(X_i - X_{i'})$ is obviously subgaussian as a sum of subgaussian random variables, however, its variance proxy could grow linearly with p_n.

Previous work has shown that a Gaussian random vector $Z \sim N(0, \Sigma)$ is “sign-subgaussian”, that is, $\operatorname{sign}(Z)$ is subgaussian with variance proxy that depends on p_n only through $C(\Sigma)$, for special cases when the covariance Σ is identity or equicorrelation matrix (Han and Liu, 2013). However, a result for general covariance structures was previously unknown.

In the following lemma, we resolve this question, proving that Gaussian vectors are sign-subgaussian:

Lemma 4.5. *Let $Z \sim N(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{p\times p}$. Then* $\operatorname{sign}(Z)$ *is* $\mathsf{C}(\Sigma)$*-subgaussian.*

This lemma is the primary tool for our main results in this paper—specifically, it is the key ingredient to the proof of Theorem 4.2, which bounds the errors $\breve{\Theta} - \widetilde{\Theta}$ and $\breve{S}_{ab} \cdot \det(\breve{\Theta}) - S_{ab} \cdot \det(\Theta)$. Lemma 4.5 is proved in Appendix A. We also use this result in establishing results in the following section.

4.3 Deterministic and probabilistic bounds on $\widehat{\widehat{\Sigma}} - \Sigma$.

Lemma 4.5 is instrumental in obtaining probabilistic bounds on $\widehat{\widehat{\Sigma}} - \Sigma$. Results given in this section are crucial for establishing Theorem 4.2 and Corollary 3.8.

Let $\mathcal{S}_k$ be the set of k-sparse vectors in the unit ball,

$$\mathcal{S}_k = \{u \in \mathbb{R}^p : ||u||_2 \leqslant 1, ||u||_0 \leqslant k\} \ .$$

The following lemma provides uniform deviation of $u^\top \widehat{T} u$ from $u^\top T u$ over $\mathcal{S}_k$, with the proof given in Appendix D.4.

Lemma 4.6. *Suppose that $k \geqslant 1$ and $\delta \in (0, 1)$ satisfy $\log(2/\delta) + k \log(12p) \leqslant n$. Then with probability at least $1 - \delta$ it holds that*

$$\sup_{u \in \mathcal{S}_k} \left| u^\top (\widehat{T} - T) u \right| \leqslant 16(1 + \sqrt{5}) \mathsf{C}(\Sigma) \cdot \sqrt{\frac{\log(2/\delta) + k \log(12p)}{n}} \ .$$

Next, we relate $\widehat{\widehat{\Sigma}}$ to $\widehat{T}$. First, for any $k \geqslant 1$, let $\mathcal{B}_k$ be the set[5]

$$\mathcal{B}_k = \left\{ u \in \mathbb{R}^p : \sqrt{||u||_2^2 + \frac{||u||_1^2}{k}} \leqslant 1 \right\} \ .$$

The intuition for this set is that it contains vectors bounded both in the ℓ_2 and ℓ_1 norms; it is a relaxation of k-sparsity. We have the following deterministic bound on the error of the covariance estimator $\widehat{\widehat{\Sigma}}$, which is proven in Appendix D.3:

Lemma 4.7. *The following bound holds deterministically: for any $k \geqslant 1$,*

$$\sup_{u,v \in \mathcal{B}_k} \left| u^\top (\widehat{\widehat{\Sigma}} - \Sigma) v \right| \leqslant \frac{\pi^2}{8} \cdot k ||\widehat{T} - T||_\infty^2 + 2\pi \sup_{u \in \mathcal{S}_{k+1}} \left| u^\top (\widehat{T} - T) u \right| \ . \tag{4.1}$$

[5] Note that $\mathcal{B}_k$ is the unit ball for the norm given by $||u||_{(k)} := \sqrt{||u||_2^2 + \frac{||u||_1^2}{k}}$.

Lemma B.2, given in Appendix B.2, bounds $||\widehat{T} - T||_\infty$ with high probability. Therefore, combining it with Lemma 4.7 and 4.6, we immediately obtain the following corollary:

Corollary 4.8. *Take any $\delta_1, \delta_2 \in (0,1)$ and any $k \geqslant 1$ such that $\log(2/\delta_2) + (k+1)\log(12p) \leqslant n$. Then, with probability at least $1 - \delta_1 - \delta_2$, the following bound on $\widehat{\Sigma} - \Sigma$ holds:*

$$\sup_{u,v \in \mathcal{B}_k} \left| u^\top (\widehat{\Sigma} - \Sigma) v \right| \leqslant \frac{\pi^2}{8} \cdot k \cdot \frac{4\log\left(2\binom{p}{2}/\delta_1\right)}{n} + 2\pi \cdot 16(1+\sqrt{5})\mathsf{C}(\Sigma) \cdot \sqrt{\frac{\log(2/\delta_2) + (k+1)\log(12p)}{n}}\,. \tag{4.2}$$

Results of Lemma 4.6 and Corollary 4.8 can be compared to Theorem 2 in Mitra and Zhang (2014). When $C(\Sigma) = O(1)$, we extend their result to the transelliptical copula model and provide an alternative proof for the Gaussian copula model. We note that their result does not depend on the condition number of the covariance matrix, but only on the maximum eigenvalue. However, in the context of graphical models it is commonly assumed that the smallest eigenvalue is a constant. Furthermore, our result can also be compared with Theorem 4.10 of Han and Liu (2013). We rigorously establish the result for all well-conditioned covariance matrices, without explicitly making the sign-subgaussian assumption.

5 Simulation studies

In this section, we illustrate finite sample properties of ROCKET described in Section 2. We use ROCKET to construct confidence intervals for edge parameters and report empirical coverage probabilities as well as the length of constructed intervals. For comparison, we also construct confidence intervals using the procedure of Ren et al. (2013), which is based on the Pearson correlation matrix, and a nonparanormal estimator of the correlation matrix proposed in Liu et al. (2009). For these two methods, we use the plugin estimate of the correlation matrix together with (2.7) to estimate Ω_{ab}. Recall that Liu et al. (2009) estimate the correlation matrix based on the marginal transformation of the observed data. Let

$$\widetilde{F}_a(x) = \begin{cases} \delta_n & \text{if } \widehat{F}_a(x) < \delta_n \\ \widehat{F}_a(x) & \text{if } \delta_n \leqslant \widehat{F}_a(x) \leqslant 1 - \delta_n \\ 1 - \delta_n & \text{if } \widehat{F}_a(x) > 1 - \delta_n, \end{cases}$$

where $\widehat{F}_a(x) = n^{-1}\sum_i \mathbb{I}\{X_{ia} < x\}$ is the empirical CDF of X_a and $\delta_n = \left(4n^{1/4}\sqrt{\pi \log(n)}\right)^{-1}$. The correlation matrix $\widehat{\Sigma} = \left(\widehat{\Sigma}_{ab}\right)_{ab}$ is then estimated as $\widehat{\Sigma}_{ab} = \widehat{\mathrm{Corr}}\left(\Phi\left(\widetilde{F}_a(X_{ia})\right), \Phi\left(\widetilde{F}_b(X_{ib})\right)\right)$ where $\Phi(\cdot)$ is a CDF of a standard normal distribution. Asymptotic variance of estimators of Ω_{ab} based on the Pearson or nonparanormal correlation matrix is obtained as $\breve{S}^2_{ab} = n^{-1}\left(\breve{\Omega}_{aa}\breve{\Omega}_{bb} + \breve{\Omega}^2_{ab}\right)$. For all simulations, we set the tuning parameter $\lambda = 2.1\sqrt{\log(p_n)/n}$, as suggested by our theory. Note that the constant in front of the parameter is chosen large enough so that the penalty dominates the variance of an element of the score. All computations for simulations and for the real data experiment are carried out in Matlab (MATLAB, 2014).

Simulation 1. We generate data from the model

$$X \sim \mathsf{E}(0, \Sigma, \xi), \tag{5.1}$$

		ROCKET		Pearson		Nonparanormal	
		Coverage (%)	Width	Coverage (%)	Width	Coverage (%)	Width
	$\omega_{10,11} = 10.38$	92.8	10.26	55.3	4.99	32.8	3.57
	$\omega_{10,12} = 0$	96.0	9.81	64.8	4.68	73.5	3.35
	$\omega_{10,20} = 0$	96.1	11.08	64.8	4.92	76.7	3.50
Oracle	$\omega_{10,11} = 10.38$	93.3	10.21	54.5	4.87	31.1	3.53
	$\omega_{10,12} = 0$	96.0	9.75	63.5	4.62	73.2	3.33
	$\omega_{10,20} = 0$	96.3	11.02	63.6	4.86	77.0	3.47

Table 1: Empirical coverage and average length of 95% confidence intervals based on 1000 independent simulation runs. Data generated from the model in (5.1) with chain graph structure.

where ξ follows a t-distribution with 5 degrees of freedom. The inverse covariance matrix Ω encodes one of the following structures:

- chain structure with $\Omega^0_{j,j+1} = \Omega^0_{j+1,j} = 0.5$, and
- a grid where each node is connected to its four nearest neighbors with the nonzero elements of Ω^0 equal to $\omega = 0.24$.

Diagonal element of Ω^0 are equal to 1. Let $\left(\Omega^0\right)^{-1} = \Sigma^0$. Then $\Sigma = \operatorname{diag}^{-1/2}\left(\Sigma^0\right)\Sigma^0\operatorname{diag}^{-1/2}\left(\Sigma^0\right)$ and $\Omega = \Sigma^{-1}$.

First, we consider a chain graph and generate $n = 400$ samples from model in (5.1) with $p = 1000$. Figures 2 and 3 show Q-Q plots based on 1000 independent realizations of the test statistic defined in (2.9), $\sqrt{n}\widecheck{\mathsf{Err}}_{ab}$, for the three methods together with the reference line showing quantiles of the standard normal distribution. First row in the two figures illustrates actual performance of the methods, while the second row illustrates performance of an oracle procedure that does not need to solve a high-dimensional variable selection problem, but instead knows the sparsity pattern of Ω. From these two figures, we observe that the quantiles of the test statistic $\sqrt{n}\widecheck{\mathsf{Err}}_{ab}$ based on ROCKET estimator are closest to quantiles of the standard normal random variable. We further quantify these results in Table 1, which reports empirical coverage of the confidence intervals based on $\sqrt{n}\widecheck{\mathsf{Err}}_{ab}$. From the table, we can observe that the coverage of the confidence intervals based on ROCKET is close to nominal coverage of 95%.

Similar results are seen in Figure 4, which is based on $n = 400$ samples generated from the model in (5.1) when Ω encodes the 30×30 grid structure ($p = 900$). Table 2 reports empirical coverage and length of the confidence intervals. These two examples are not surprising, since neither the Pearson nor the nonparanormal correlation matrix consistently estimate the true Σ under the model in (5.1). However, using ROCKET we are able to construct a test statistic $\sqrt{n}\widecheck{\mathsf{Err}}_{ab}$ that is asymptotically distributed as a standard normal random variable. The asymptotic distribution provides a good approximation to the finite sample distribution of $\sqrt{n}\widecheck{\mathsf{Err}}_{ab}$. We also observe that ROCKET performs similarly to the oracle procedure that knows the sparsity structure of Ω. Note

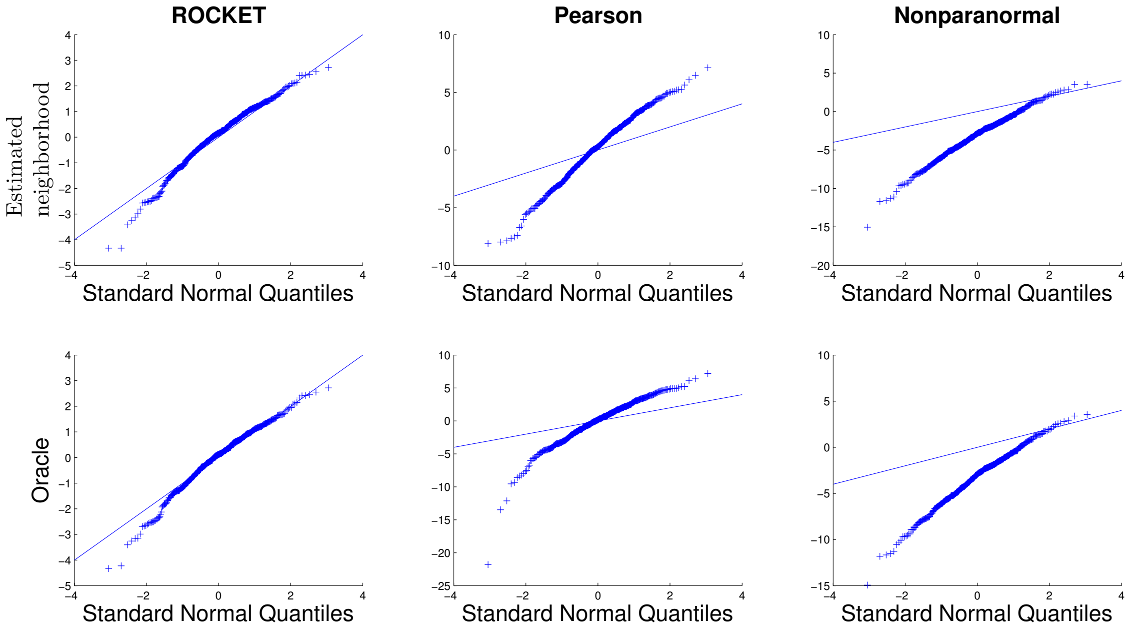

Figure 2: Simulation 1. Q-Q plot of $\sqrt{n}\breve{\mathsf{Err}}_{ab}$ with $a = 10$ and $b = 11$ (edge) when data are generated from the model in (5.1) with Ω encoding the chain structure.

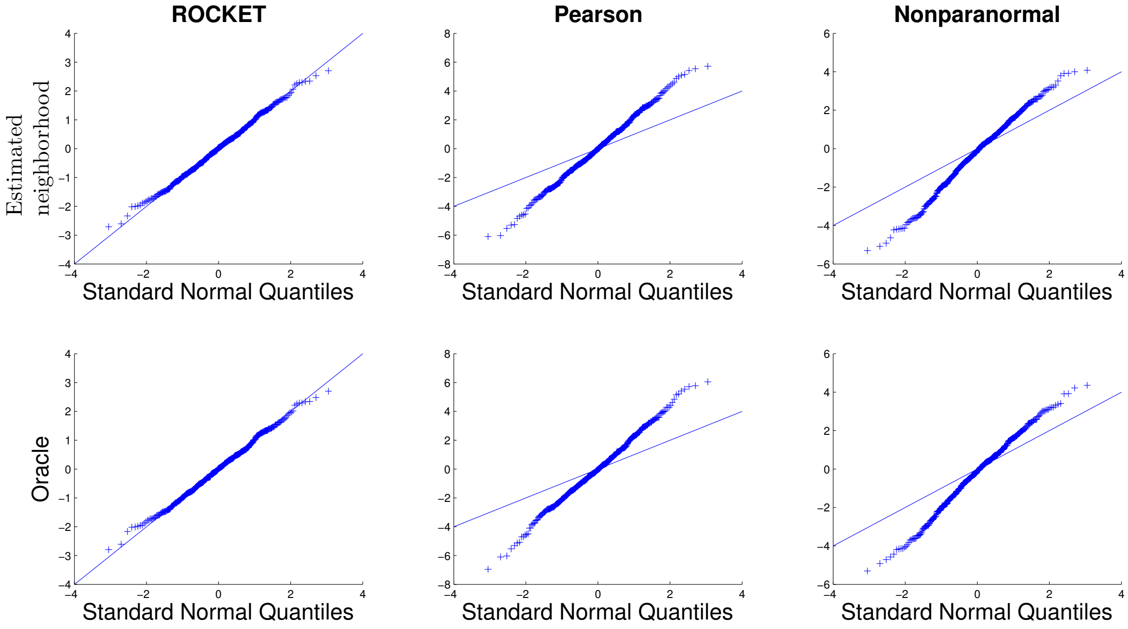

Figure 3: Simulation 1. Q-Q plot of $\sqrt{n}\breve{\mathsf{Err}}_{ab}$ with $a = 10$ and $b = 12$ (non-edge close to an edge) when data are generated from the model in (5.1) with Ω encoding the chain structure.

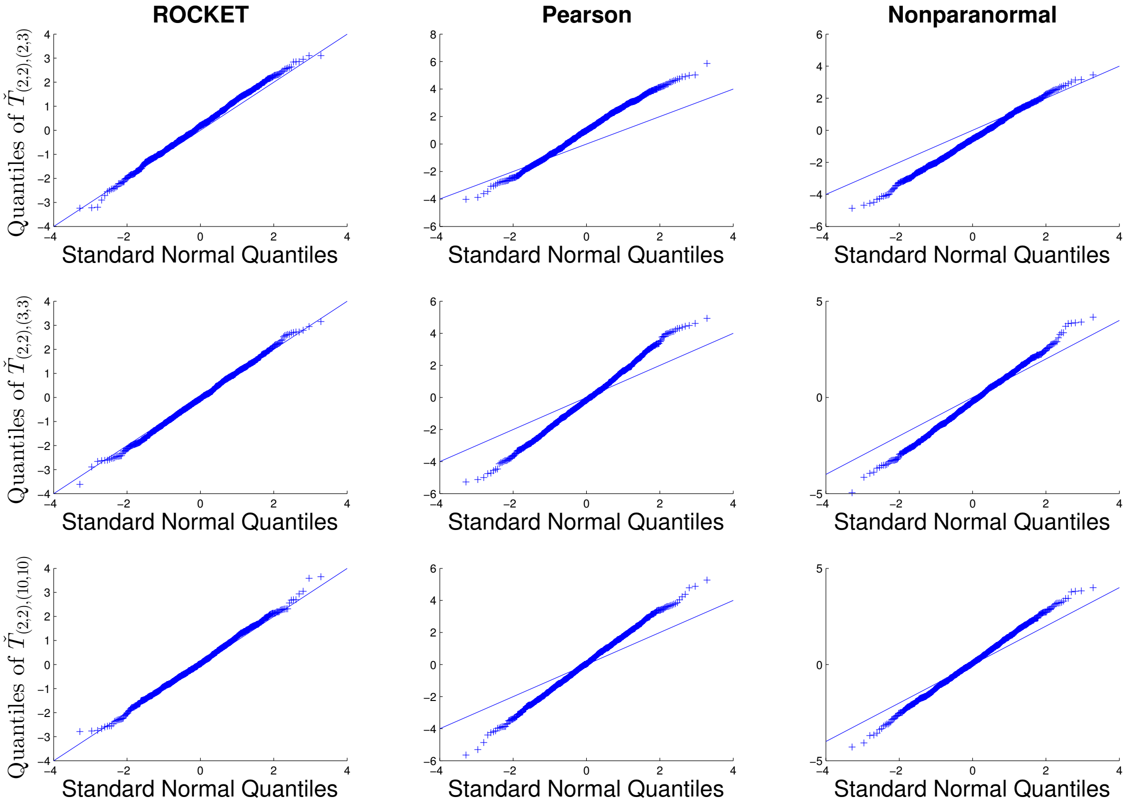

Figure 4: Simulation 1. Q-Q plot of $\sqrt{n}\breve{\mathsf{Err}}_{ab}$ when data are generated from the model in (5.1) with Ω encoding the grid structure. First row corresponds to an edge, second row to a close non-edge, and third row to a far non-edge.

that the width of the confidence intervals obtained from ROCKET is larger due to imperfect model selection.

Simulation 2. We illustrate performance of ROCKET when data are generated from a normal and nonparanormal distribution. We consider Ω corresponding to a grid described in Simulation 1 and generate $n = 400$ samples from $N(0, \Omega^{-1})$ and $\mathsf{NPN}(\Omega^{-1}; \widetilde{f}_1, \ldots, \widetilde{f}_p)$, where $\widetilde{f}_j = f_{\mathrm{mod}(j-1,5)+1}$ with $f_1(x) = x$, $f_2(x) = \mathrm{sign}(x)\sqrt{|x|}$, $f_3(x) = x^3$, $f_4(x) = \Phi(x)$, and $f_5(x) = \exp(x)$. Here $\mathrm{mod}(a, b)$ denotes the remainder after division of a by b.

Table 3 summarizes results from the simulation. We observe that when data are multivariate normal all three methods perform well, with ROCKET having slightly wider intervals, but with similar coverage. When data are generated from a nonparanormal distribution, using the Pearson correlation in (2.7) results in confidence intervals that do not have nominal coverage due to the bias. In this setting, nonparanormal estimator and ROCKET still have the correct nominal coverage. Note however that when Kendall's tau is equal to zero, Pearson correlation is also equal to zero. See, for example, coverage for $\omega_{(2,2),(3,3)}$ and $\omega_{(2,2),(10,10)}$.

Similar results were obtained when data are generated from $\mathsf{TE}(\Omega^{-1}, \xi; \widetilde{f}_1, \ldots, \widetilde{f}_p)$ with $\xi \sim t_1$ or $\xi \sim t_5$. Due to space constraints, results are not shown.

	ROCKET		Pearson		Nonparanormal	
	Coverage (%)	Width	Coverage (%)	Width	Coverage (%)	Width
$\omega_{(2,2),(2,3)} = 0.41$	92.8	0.54	66.6	0.49	79.7	0.34
$\omega_{(2,2),(3,3)} = 0$	93.5	0.56	74.2	0.47	82.8	0.33
$\omega_{(2,2),(10,10)} = 0$	93.8	0.57	74.8	0.48	85.3	0.33

Table 2: Simulation 1. Empirical coverage and average length of 95% confidence intervals based on 1000 independent simulation runs. Data generated from the model in (5.1) with grid graph structure.

		ROCKET		Pearson		Nonparanormal	
		Coverage (%)	Width	Coverage (%)	Width	Coverage (%)	Width
Gaussian	$\omega_{(2,2),(2,3)} = 0.41$	93.3	0.37	93.3	0.35	93.3	0.35
	$\omega_{(2,2),(3,3)} = 0$	94.7	0.38	94.1	0.34	93.9	0.34
	$\omega_{(2,2),(10,10)} = 0$	94.7	0.38	95.2	0.34	95.2	0.34
Transformed Gaussian	$\omega_{(2,2),(2,3)} = 0.41$	93.4	0.37	0.0	0.26	94.8	0.35
	$\omega_{(2,2),(3,3)} = 0$	94.9	0.38	89.4	0.29	95.5	0.34
	$\omega_{(2,2),(10,10)} = 0$	94.7	0.38	95.3	0.28	94.4	0.34

Table 3: Simulation 2. Empirical coverage and average length of 95% confidence intervals based on 1000 independent simulation runs. Ω corresponds to a grid graph structure.

Simulation 3. In this simulation, we illustrate the power of a test based on $\sqrt{n}\breve{\mathsf{Err}}_{ab}$ to reject the null hypothesis $H_{0,ab} : \Omega_{ab} = 0$. Samples are generated from the model in (5.1) with ξ having χ_p, t_5, and t_1 distribution and the covariance matrix is of the form $\Sigma = I_P + E$ where $E_{12} = E_{21} = \rho$ with $p = 1000$ and $n = 400$. Note that $\xi \sim \chi_p$ implies that X is multivariate normal. We also consider marginal transformation of X as described in Simulation 2. Figure 5 plots empirical power curves based on 1000 independent simulation runs for different settings. When data are following normal distribution all three methods have similar power. For other distributions, tests based on Pearson and nonparanormal correlation do not have correct coverage and are shown for illustrative purpose only.

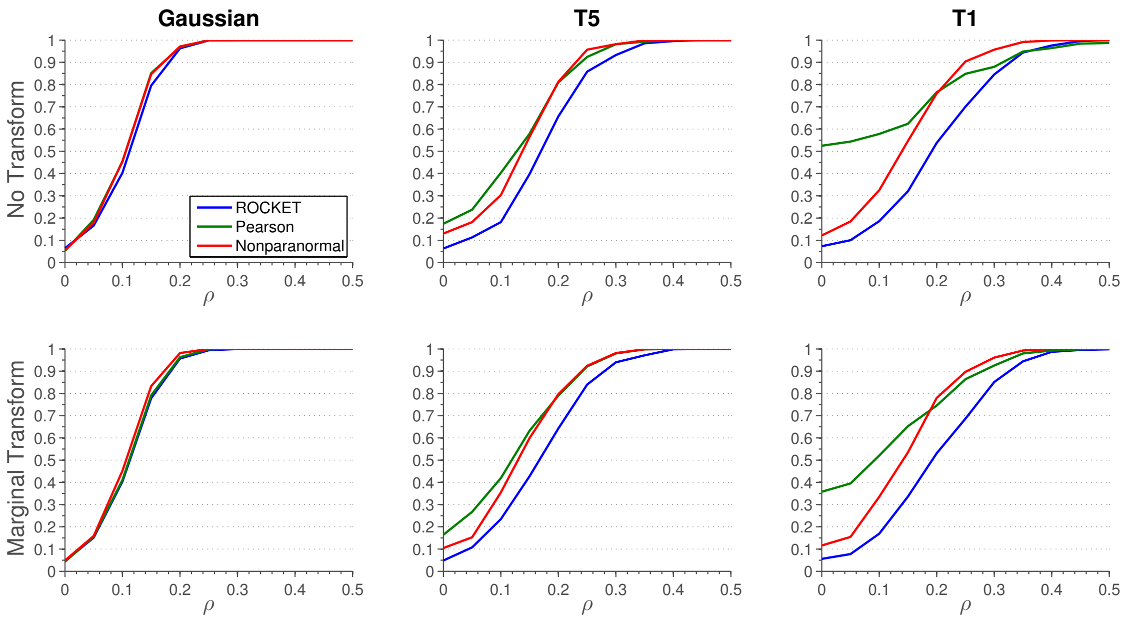

Figure 5: Simulation 3. Power plots for simulated data generated from a Gaussian distribution, and from a multivariate t distribution with 5 degrees of freedom and with 1 degree of freedom.

6 Real data experiment

In this section, we evaluate the performance of the ROCKET method on a real data set, and compare with the Gaussian graphical model based approach of Ren et al. (2013) (using Pearson correlation) and the nonparanormal estimator proposed in Liu et al. (2009) (details for these methods are given in Section 5).

We use stock price closing data obtained via the R package `huge` (Zhao et al., 2014), which was gathered from publicly available data from Yahoo Finance[6]. The data consists of daily closing prices of 452 S&P 500 companies over 1258 days. We transform the data to consider the log-returns, that is, we form a matrix $X \in \mathbb{R}^{1258\times 452}$ with entries

$$X_{ij} = \log\left(\frac{\text{Closing price of stock } j \text{ on day } i+1}{\text{Closing price of stock } j \text{ on day } i}\right) .$$

While in practice there is dependence across time in this data set, we treat each row of X as independent.

We perform two experiments on this data set. In Experiment 1, we test whether empirical results agree with the asymptotic normality predicted by the theory for the three methods—we do this by splittting the data into disjoint subsamples and comparing estimates across these subsamples. In Experiment 2, we use the full sample size and compare the estimates and confidence intervals produced by each of the three methods.

[6] http://ichart.finance.yahoo.com

6.1 Experiment 1: checking asymptotic normality

In this real data example, there is no available "ground truth" to compare to—that is, we do not know the true distribution of the data, and cannot compare our estimates to an exact true precision matrix Ω. However, we can still check whether the estimators produced by these methods exhibit asymptotic normality (as claimed in the theory), by splitting the data into many subsamples and considering the empirical distribution of the estimators across these subsamples.

We will split the data into $L = 25$ subsamples of size $n = 50$ each. Due to this small sample size, we restrict our attention to companies in the categories `Materials` and `Consumer Staples`, which consist of 29 and 35 companies, respectively, for a total of $p = 64$ companies. To construct our subsampled data, we randomly select $L = 25$ disjoint sets of size $n = 50$ from $\{1, \ldots, 1257\}$, denoted as $I_1, \ldots, I_L$. For each $\ell = 1, \ldots, 25$, define the ℓth data set

$$X^{(\ell)} = X_{I_\ell, S} \in \mathbb{R}^{n \times p} ,$$

where $S \subset \{1, \ldots, 452\}$ identifies the $p = 64$ stocks of interest.

Next, for each pair (a, b) of stocks, with $1 \leqslant a < b \leqslant p$, and for each subsample ℓ, we compute $\breve{\Omega}^{(\ell)}_{ab}$ and $\breve{S}^{(\ell)}_{ab}$ using the ROCKET method. Suppose that the true distribution of the data follows the transelliptical model with precision matrix Ω. Recall that our main result, Theorem 3.5, implies that $\sqrt{n} \cdot \frac{\breve{\Omega}^{(\ell)}_{ab} - \Omega_{ab}}{\breve{S}^{(\ell)}_{ab}}$ is approximately distributed as a standard normal. Since $\breve{S}^{(\ell)}_{ab}$ concentrates near S_{ab} (the asymptotic variance calculated in Theorem 4.1), we see that we should have

$$z^{(\ell)}_{ab} := \sqrt{n} \cdot \frac{\breve{\Omega}^{(\ell)}_{ab}}{\breve{S}^{(\ell)}_{ab}} \approx \sqrt{n} \cdot \frac{\Omega_{ab}}{S_{ab}} + N(0, 1) .$$

Therefore, writing $\mu_{ab} := \sqrt{n} \cdot \Omega_{ab}/S_{ab}$, we should have

$$(z^{(1)}_{ab}, \ldots, z^{(L)}_{ab}) \approx \mu \cdot \mathbf{1}_L + N(0, \mathbf{I}_L) .$$

In particular, this implies that the sample variance of this vector should have expectation

$$\mathsf{SampleVar}(z_{ab}) \mathbb{E} \frac{1}{L-1} \sum_{\ell=1}^{L} \left(z^{(\ell)}_{ab} - \overline{z}_{ab} \right)^2 \approx 1 ,$$

where $\overline{z}_{ab} := \frac{1}{L} \sum_\ell z^{(\ell)}_{ab}$.

In Figure 6, we show a histogram of the sample variances $\mathsf{SampleVar}(z_{ab})$ across all $\binom{p}{2} = 2016$ pairs of variables. To compare to the Pearson and nonparanormal methods, we repeat this procedure for the estimators (and estimated variances) produced by the other two methods as well, which are also displayed in Figure 6. We see that ROCKET produces a mean sample variance ≈ 0.98 (very near to 1), while the other two methods give mean sample variances of ≈ 1.28 (Pearson) and ≈ 1.265 (nonparanormal), substantially higher than the theoretical value of 1. This indicates that the normal approximation to the distribution of the estimator may be approximately valid for ROCKET, but does not have the correct scale (that is, the scale predicted by the theory) for the other two methods, on this data set.

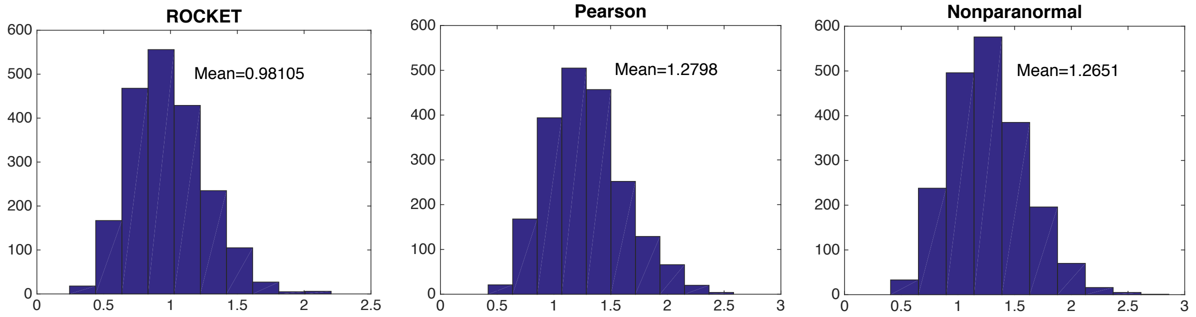

Figure 6: Sample variances of the rescaled estimator, $\sqrt{n} \cdot \check{\Omega}_{ab}/\check{S}_{ab}$, for each pair of variables (a, b), using the subsampled stock data. The sample variances should be approximately 1 according to the theory (see Section 6.1).

The vector $(z_{ab}^{(1)}, \dots, z_{ab}^{(L)})$, in addition to having sample variance near 1, should also exhibit Gaussian-like tails according to the theory. To check this, we calculate the proportion of values in this vector lying near to the mean, specifically,

$$\frac{\#\left\{\ell : \left|z_{ab}^{(\ell)} - \overline{z}_{ab}\right| \leqslant 1.6449\sqrt{1 - \frac{1}{L}}\right\}}{L},$$

which should be approximately 90% according to the theory (since the standard normal distribution has 90% of its mass between ±1.6449). The results are:

Method	Coverage (theory: 90%)
ROCKET	90.55%
Pearson	85.01%
Nonparanormal	85.18%

We see that only the ROCKET method achieves the appropriate coverage.

6.2 Experiment 2: estimating a graph

In the second experiment, we use the full sample size $n = 1257$ to estimate a sparse graph over the $p = 64$ stocks selected for Experiment 1, using each of the three methods. To do this, for each method we first produce a (approximate) p-value testing for the presence of an edge between each pair (a, b) of variables. Recall that according to our main result, Theorem 3.5, if the pair of variables (a, b) does not have an edge, then $\Omega_{ab} = 0$ and so $\sqrt{n} \cdot \check{\Omega}_{ab}/\check{S}_{ab}$ is approximately distributed as a standard normal variable. Then, using a two-sided z-test, we calculate a p-value

$$p_{ab} = 2 - 2\Phi\left(\left|\sqrt{n} \cdot \check{\Omega}_{ab}/\check{S}_{ab}\right|\right) .$$

In Figure 7, we show the resulting graphs when edge (a, b) is drawn whenever the p-value passes the threshold $p_{ab} < 0.00001$ or whenever $p_{ab} < 0.001$. The number of edges selected for each method is shown in the figures. Overall we see that ROCKET selects roughly the same number of edges as the Pearson method but less than the nonparanormal method, on this data set. To further

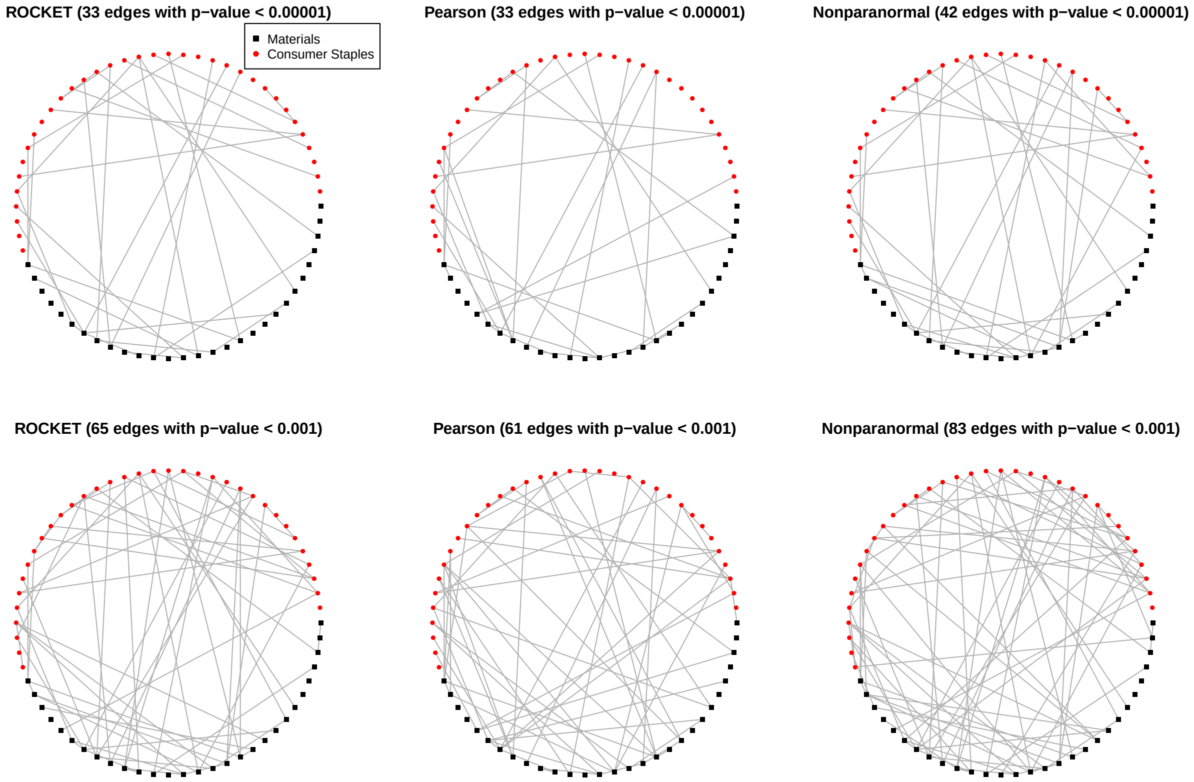

Figure 7: Estimated graph for the stock data, using the ROCKET, Pearson, and nonparanormal methods (see Section 6.2). An edge is displayed for each pair of variables (a, b) with p-value $p_{ab} < 0.00001$ (top row) and $p_{ab} < 0.001$ (bottom row). Graphs were drawn using the `igraph` package (Csardi and Nepusz, 2006) in R (R Core Team, 2012).

compare the methods, in Figure 8 we show the distribution of the p-values p_{ab} across all pairs of variables (a, b), for each method. ROCKET produces slightly less low (strong) p-values than the nonparanormal method, and slightly more low (strong) p-values than the Pearson method, on this data set.

Since the Pearson and nonparanormal methods do not exhibit approximately normal behavior across subsamples (Experiment 1), this should not be interpreted as a power comparison between the methods; the additional edges selected by the nonparanormal method, for instance, may not be as reliable since the p-value calculation is based on approximating the distribution of the estimator using a theoretical scaling that does not appear to hold for this method.

7 Discussion

We have proposed a novel procedure ROCKET for inference on elements of the latent inverse correlation matrix under high-dimensional elliptical copula models. Our paper has established a surprising result, which states that ROCKET produces an asymptotically normal estimator for

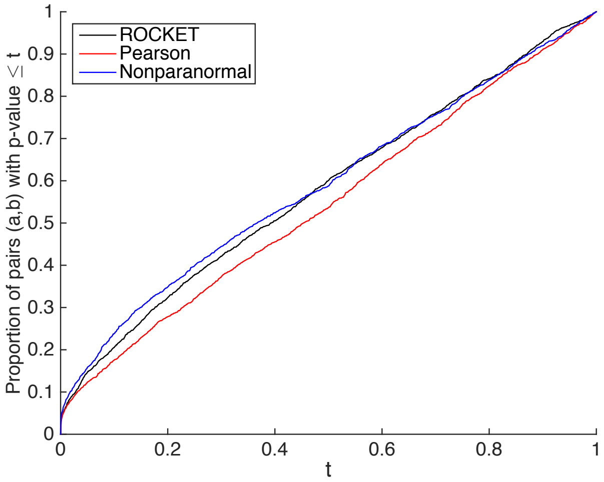

Figure 8: Distribution of p-values p_{ab} across all pairs of variables (a, b) on the stock data, for the ROCKET, Pearson, and nonparanormal methods (see Section 6.2).

an element of the inverse correlation matrix in an elliptical copula model with the same sample complexity that is required to obtain an asymptotically normal estimator for an element in the precision matrix under a multivariate normal distribution. Furthermore, this sample complexity is optimal (Ren et al., 2013). The result is surprising as the family of elliptical copula models is much larger than the family of multivariate normal distributions. For example, it contains distributions with heavy tail dependence as discussed in Section 2. ROCKET achieves the optimal requirement on the sample size without knowledge of the marginal transformation. Our result is also of significant practical importance. Since normal distribution is only a convenient mathematical approximation to data generating process, we recommend using ROCKET whenever making inference about inverse correlation matrix, instead of methods that heavily rely on Normality. From simulation studies, even when data are generated from a normal distribution, ROCKET does not lose power compared to procedures that were specifically developed for inference under Normality.

The main technical tool developed in the paper establishes that the sign of normal random vector, taken elementwise, is itself a sub-Gaussian random variable with the sub-Gaussian parameter depending on the condition number of the covariance matrix Σ. Based on this result, we were able to establish a tight tail bound on the deviation of sparse eigenvalues of the Kendall's tau matrix $\widehat{T}$. This result is of independent interest and it would allow us to improve a number of recent results on sparse principal component analysis, factor models and estimation of structured covariance matrices (Mitra and Zhang, 2014; Han and Liu, 2013; Fan et al., 2014). The sharpest result on the nonparametric estimation of correlation matrices in spectral norm under a Gaussian copula model was established in (Mitra and Zhang, 2014). We extend its validity to the family of elliptical copula models and provide an alternative proof. Previous work of Han and Liu (2013) and Fan et al. (2014) has established sub-Guassianity of the sign vector for special cases of covariance matrices (identity and equi-correlation matrix). Our work rigorously proves the result for the class of well-conditioned covariance matrices.

Acknowledgments

This work is partially supported by an IBM Corporation Faculty Research Fund at the University of Chicago Booth School of Business. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

References

R. G. Baraniuk, M. A. Davenport, , and M. B. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli‘*, 19(2):521–547, 2013.

A. Belloni, V. Chernozhukov, and C. B. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.*, 81(2):608–650, 2013a.

A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv preprint arXiv:1312.7186*, 2013b.

A. Belloni, V. Chernozhukov, and K. Kato. Uniform post selection inference for lad regression models. *arXiv preprint arXiv:1304.0282*, 2013c.

A. Belloni, V. Chernozhukov, and Y. Wei. Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969*, 2013d.

P. Bühlmann and S. A. van de Geer. *Statistics for high-dimensional data.* Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

T. T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

H. Callaert and P. Janssen. The Berry-Esseen theorem for U-statistics. *Ann. Stat.*, 6(2):417–421, 1978.

E. J. Candés and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.*, 35(6):2313–2351, 2007.

M. Chen, Z. Ren, H. Zhao, and H. H. Zhou. Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *arXiv preprint arXiv:1309.5923*, 2013.

J. Cheng, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *ArXiv e-prints, arXiv:1304.2810*, 2013.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30(1):56–66, 2008.

V. de la Pena and E. Giné. *Decoupling: from dependence to independence.* Springer, 1999.

P. Embrechts, F. Lindskog, and A. McNeil. Modelling dependence with copulas and applications to risk management. In S. T. Rachev, editor, *Handbook of heavy tailed distributions in finance*, pages 329–384. Elsevier, 2003.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *Ann. Appl. Stat.*, 3(2):521–541, 2009.

J. Fan, F. Han, and H. Liu. Page: Robust pattern guided estimation of large covariance matrix. Technical report, Technical report, Princeton University, 2014.

K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*, volume 36 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, Ltd., London, 1990.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint arXiv:1309.4686*, 2013.

J. H. Friedman, T. J. Hastie, and R. J. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Q. Gu, Y. Cao, Y. Ning, and H. Liu. Local and global inference for high dimensional gaussian copula graphical models. *ArXiv e-prints, arXiv:1502.02347*, 2015.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011a.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Asymptotic properties of the joint neighborhood selection method for estimating categorical markov networks. Technical report, University of Michigan, 2011b.

F. Han and H. Liu. Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *ArXiv e-prints, arXiv:1305.6916*, 2013.

H. Höfling and R. J. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.*, 10:883–906, 2009.

A. Javanmard and A. Montanari. Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274*, 2013.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(Oct):2869–2909, 2014.

C. Klüppelberg, G. Kuhn, and L. Peng. Semi-parametric models for the multivariate tail dependence function–the asymptotically dependent case. *Scand. J. Stat.*, 35(4):701–718, 2008.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.*, 37:4254–4278, 2009.

S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

J. D. Lee and T. J. Hastie. Learning mixed graphical models. *ArXiv e-prints, arXiv:1205.5012*, 2012.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference with the lasso. *ArXiv e-prints, arXiv:1311.6238*, 2013.

F. Lindskog, A. McNeil, and U. Schmock. Kendall's tau for elliptical distributions. *Credit Risk*, pages 149–156, 2003.

H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating gaussian graphical models. *ArXiv e-prints, arXiv:1209.2437*, 2012.

H. Liu, J. D. Lafferty, and L. A. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

H. Liu, F. Han, M. Yuan, J. D. Lafferty, and L. A. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.*, 40(4):2293–2326, 2012a.

H. Liu, F. Han, and C.-H. Zhang. Transelliptical graphical models. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Proc. of NIPS*, pages 809–817. 2012b.

W. Liu. Gaussian graphical model estimation with false discovery rate control. *Ann. Stat.*, 41(6): 2948–2978, 2013.

R. Lockhart, J. E. Taylor, R. J. Tibshirani, and R. J. Tibshirani. A significance test for the lasso. *Ann. Stat.*, 42(2):413–468, 2014.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv preprint arXiv:1305.2436*, 2013.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

MATLAB. *version 8.4.0 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Stat.*, 34(3):1436–1462, 2006.

R. Mitra and C.-H. Zhang. Multivariate analysis of nonparametric estimates of large correlation matrices. *ArXiv e-prints, arXiv:1403.6195*, 2014.

R. I. Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. *ArXiv e-prints, arXiv:1312.2903*, 2013.

T. Peel, S. Anthoine, and L. Ralaivola. Empirical bernstein inequalities for u-statistics. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Adv. Neural Inf. Process. Syst. 23*, pages 1903–1911. Curran Associates, Inc., 2010.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319, 2010.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.

A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Stat.*, 2:494–515, 2008.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Learning theory*, volume 3559 of *Lecture Notes in Comput. Sci.*, pages 545–560. Springer, Berlin, 2005.

J. E. Taylor, R. Lockhart, R. J. Tibshirani, and R. J. Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 2014.

S. A. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.*, 42(3):1166–1202, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications.* Cambridge University Press, 2012.

M. Wegkamp and Y. Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *ArXiv e-prints, arXiv:1305.6526*, 2013.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Stat.*, 40(5):2541–2571, 2012.

L. Xue, H. Zou, and T. Ca. Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *Ann. Stat.*, 40(3):1403–1429, 2012.

E. Yang, G. I. Allen, Z. Liu, and P. Ravikumar. Graphical models via generalized linear models. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1358–1366. Curran Associates, Inc., 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On graphical models via univariate exponential family distributions. *ArXiv e-prints, arXiv:1301.4183*, 2013.

E. Yang, Y. Baker, P. Ravikumar, G. I. Allen, and Z. Liu. Mixed graphical models via exponential families. In *Proc. 17th Int. Conf, Artif. Intel. Stat.*, pages 1042–1050, 2014.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286, 2010.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B*, 76(1):217–242, 2013.

T. Zhao and H. Liu. Calibrated precision matrix estimation for high dimensional elliptical distributions. *IEEE Trans. Inf. Theory*, pages 1–1, 2014.

T. Zhao, H. Liu, K. Roeder, J. D. Lafferty, and L. A. Wasserman. *huge: High-dimensional Undirected Graph Estimation*, 2014. R package version 1.2.6.

A Gaussian vectors and the sign-subgaussian property

In this section we prove Lemma 4.5, which shows that that a centered Gaussian vector $z \sim N(0, \Sigma)$ satisfies the sign-subgaussianity property, that is, the sign vector $\mathrm{sign}(z)$ is itself subgaussian.

Lemma 4.5. *Let* $Z \sim N(0, \Sigma)$ *for some* $\Sigma \in \mathbb{R}^{p \times p}$*. Then* $\mathrm{sign}(Z)$ *is* $\mathsf{C}(\Sigma)$*-subgaussian.*

Proof of Lemma 4.5. Without loss of generality, rescale so that $\lambda_{\min}(\Sigma) = 1$ and then $\mathsf{C}(\Sigma) = \lambda_{\max}(\Sigma)$. Write $\Sigma = AA^\top + \mathbf{I}_p$ for some matrix $A \in \mathbb{R}^{n \times n}$. Then we can write $Z = X + AY$, where $X, Y \overset{iid}{\sim} N(0, \mathbf{I}_p)$. Then, for any fixed vector $v \in \mathbb{R}^p$,

$$\mathbb{E}\left[e^{v^\top \mathrm{sign}(Z)}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{v^\top \mathrm{sign}(Z)} \mid Y\right]\right] = \mathbb{E}\left[\mathbb{E}\left[e^{v^\top \mathrm{sign}(X+AY)} \mid Y\right]\right]$$
$$= \mathbb{E}\left[\prod_i \mathbb{E}\left[e^{v_i \,\mathrm{sign}(X_i + (AY)_i)} \mid Y\right]\right] ,$$

where the last step holds because, conditional on Y, each of the terms $\mathrm{sign}(X_i + (AY)_i)$ depends on X_i only, and therefore these terms are conditionally independent. Next, observe that

$$\mathbb{E}\left[\mathrm{sign}(X_i + (AY)_i) \mid Y\right] = \mathbb{E}\left[\mathrm{sign}\left(N(0,1) + (AY)_i\right) \mid Y\right] = \Phi\left((AY)_i\right) - \Phi\left(-(AY)_i\right) = \psi\left((AY)_i\right) ,$$

where we define $\psi(z) = \Phi(z) - \Phi(-z)$ for $z \in \mathbb{R}$. Then, for each i,

$$\mathbb{E}\left[e^{v_i \,\mathrm{sign}(X_i + (AY)_i)} \mid Y\right] = \mathbb{E}\left[e^{v_i (\mathrm{sign}(X_i + (AY)_i) - \psi((AY)_i))} \mid Y\right] \cdot e^{v_i \psi((AY)_i)} \leqslant e^{v_i^2/2} \cdot e^{v_i \psi((AY)_i)} ,$$

where the inequality is proved by applying Hoeffding's Lemma (see, for example, Massart (2007, Lemma 2.6)) to the bounded mean-zero random variable $[v_i\left(\mathrm{sign}(X_i + (AY)_i) - \psi\left((AY)_i\right)\right)]$. Combining the calculations so far, we have

$$\mathbb{E}\left[e^{v^\top \mathrm{sign}(Z)}\right] \leqslant e^{||v||_2^2/2} \cdot \mathbb{E}\left[e^{v^\top \psi(AY)}\right] ,$$

where $\psi(AY)$ applies the function $\psi(\cdot)$ elementwise to the vector AY.

Next we show that $y \mapsto v^\top \psi(Ay)$ is Lipschitz over $y \in \mathbb{R}^n$. Note that $x \mapsto \psi(x)$ is 1-Lipschitz over $x \in \mathbb{R}$ since the density of the standard normal distribution is bounded uniformly as $\phi(x) \leqslant \frac{1}{\sqrt{2\pi}} \leqslant \frac{1}{2}$. For any $y, y' \in \mathbb{R}^n$, we have

$$\left|v^\top \psi(Ay) - v^\top \psi(Ay')\right| \leqslant \sum_i |v_i| \cdot \left|\psi\left((Ay)_i\right) - \psi\left((Ay')_i\right)\right| \leqslant \sum_i |v_i| \cdot \left|(Ay)_i - (Ay')_i\right|$$
$$\leqslant ||v||_2 \cdot ||A(y-y')||_2 \leqslant ||v||_2 \cdot \sqrt{\lambda_{\max}(\Sigma) - 1} \cdot ||y - y'||_2 \ ,$$

where the last step is true because

$$||A||_{\mathsf{op}} = \sqrt{||\Sigma - \mathbf{I}_p||_{\mathsf{op}}} = \sqrt{\mathsf{C}(\Sigma) - 1} \ .$$

Therefore, $y \mapsto v^\top \psi(Ay)$ is $\left(||v||_2 \cdot \sqrt{||\Sigma - \mathbf{I}_p||_{\mathsf{op}}}\right)$-Lipschitz in Y. Furthermore, $\psi(x) = -\psi(-x)$ for all $x \in \mathbb{R}$, and so for any $y \in \mathbb{R}^n$,

$$v^\top \psi(Ay) = -v^\top \psi(A \cdot (-y)) \quad \Rightarrow \quad \mathbb{E}\left[\psi(AY)\right] = 0 \text{ since } Y \overset{\mathcal{D}}{=} -Y \ .$$

We can now apply standard concentration results for Lipschitz functions of a Gaussian: by Massart (2007, Proposition 3.5), $\mathbb{E}\left[e^{v^\top \psi(AY)}\right] \leqslant e^{||v||_2^2(\mathsf{C}(\Sigma)-1)/2}$. Therefore,

$$\mathbb{E}\left[e^{v^\top \operatorname{sign}(Z)}\right] \leqslant \mathbb{E}\left[e^{||v||_2^2/2 + v^\top \psi(AY)}\right] \leqslant e^{||v||_2^2/2 + ||v||_2^2(\mathsf{C}(\Sigma)-1)/2} = e^{||v||_2^2 \cdot \mathsf{C}(\Sigma)/2} \ .$$

□

B Proof of main result

B.1 Preliminaries

We first compute bounds on $||\gamma_c||_2$ and $||\gamma_c||_1$ for each $c = a, b$, which we will use many times in the proofs below. First, for $c = a, b$ note that

$$||\gamma_c||_2 = ||\Sigma_I^{-1}\Sigma_{Ic}||_2 \leqslant ||\Sigma_I^{-1}|| \cdot ||\Sigma_{Ic}||_2 \leqslant [\lambda_{\min}(\Sigma)]^{-1} \cdot \lambda_{\max}(\Sigma) \leqslant C_{\mathsf{cov}} \tag{B.1}$$

by Assumption 3.1. Next,

$$||\gamma_c||_1 = ||\Sigma_I^{-1}\Sigma_{Ic}||_1.$$

By matrix blockwise inversion,

$$= || - \Omega_{I,ab}\Theta_{ab,c}||_1 = \sum_{j \in I} |\Omega_{j,ab}\Theta_{ab,c}| \leqslant \sum_{j \in I} ||\Omega_{j,ab}||_1 ||\Theta_{ab,c}||_\infty.$$

Since $||\Theta||_\infty \leqslant \lambda_{\max}(\Theta) = \left(\lambda_{\min}(\Omega_{ab,ab})\right)^{-1} \leqslant \left(\lambda_{\min}(\Omega)\right)^{-1} = \lambda_{\max}(\Sigma) \leqslant \mathsf{C}_{\mathsf{cov}}$,

$$\leqslant C_{\mathsf{cov}} \sum_{j \in I} ||\Omega_{j,ab}||_1 = C_{\mathsf{cov}}\left(||\Omega_a||_1 + ||\Omega_b||_1\right).$$

Applying Assumption 3.2,

$$\leqslant 2C_{\mathsf{cov}}C_{\mathsf{sparse}}\sqrt{k_n} \ . \tag{B.2}$$

B.2 Proof of Theorem 4.1: asymptotic normality of the oracle estimator

Theorem 4.1. *Suppose that Assumptions 3.1, 3.2, and 3.4 hold. Then there exist constants $C_{\mathsf{normal}}, C_{\mathsf{variance}}$ depending on $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{kernel}}$ but not on (n, p_n, k_n), such that*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\sqrt{n}\cdot\frac{\widetilde{\Theta}_{ab}-\Theta_{ab}}{S_{ab}\cdot\det(\Theta)}\leqslant t\right\}-\Phi(t)\right|\leqslant C_{\mathsf{normal}}\cdot\frac{k_n\log(p_n)}{\sqrt{n}}+\frac{1}{2p_n}\ ,$$

where S_{ab} is defined in the proof and satisfies $S_{ab}\cdot\det(\Theta)\geqslant C_{\mathsf{variance}}>0$.

Proof of Theorem 4.1. We first show that the error $\widetilde{\Theta}_{ab}-\Theta_{ab}$ can be approximated by a linear function of the Kendall's tau estimator $\widehat{T}$. Define vectors $u, v\in\mathbb{R}^{p_n}$ with entries

$$u_a=1, u_b=0, u_I=-\gamma_a \text{ and } v_a=0, v_b=1, v_I=-\gamma_b\ .$$

Then by definition, we have $\widetilde{\Theta}_{ab}=u^\top\widehat{\Sigma}v$ and $\Theta_{ab}=u^\top\Sigma v$, that is, the error is given by

$$\widetilde{\Theta}_{ab}-\Theta_{ab}=u^\top(\widehat{\Sigma}-\Sigma)v\ .$$

Next, since $\widehat{\Sigma}=\sin\left(\frac{\pi}{2}\widehat{T}\right)$ and $\Sigma=\sin\left(\frac{\pi}{2}T\right)$, we take a second-order Taylor expansion of $\sin(\cdot)$ to see that, for some $t\in[0,1]$,

$$\widetilde{\Theta}_{ab}-\Theta_{ab}=u^\top\Bigg[\frac{\pi}{2}\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)- \\ \frac{1}{2}\cdot\left(\frac{\pi}{2}\right)^2\cdot\sin\left(\frac{\pi}{2}(t\cdot T+(1-t)\cdot\widehat{T})\right)\circ(\widehat{T}-T)\circ(\widehat{T}-T)\Bigg]v\ . \tag{B.3}$$

Next, we rewrite this linear term. We have

$$L:=u^\top\left[\cos\left(\frac{\pi}{2}T\right)\circ\widehat{T}\right]v=\frac{1}{\binom{n}{2}}\sum_{i<i'}\operatorname{sign}(X_i-X_{i'})^\top\left(uv^\top\circ\cos\left(\frac{\pi}{2}T\right)\right)\operatorname{sign}(X_i-X_{i'})\ ,$$

which is a U-statistic of order 2 with respect to the data $(X_1,\ldots,X_n)$. Note that

$$L-\mathbb{E}[L]=u^\top\left[\frac{\pi}{2}\cos\left(\frac{\pi}{2}T\right)\circ\widehat{T}\right]v-u^\top\left[\frac{\pi}{2}\cos\left(\frac{\pi}{2}T\right)\circ\mathbb{E}[\widehat{T}]\right]v=u^\top\left[\frac{\pi}{2}\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)\right]v\ .$$

Define the kernel $g(X,X')=\operatorname{sign}(X-X')^\top\left(uv^\top\circ\cos\left(\frac{\pi}{2}T\right)\right)\operatorname{sign}(X-X')$, and let $g_1(X)=\mathbb{E}\left[g(X,X')\mid X\right]$, where $X,X'\overset{iid}{\sim}\mathsf{TE}(\Sigma,\xi;f_1,\ldots,f_p)$. Let $\nu_{g_1}^2=\mathsf{Var}(g_1(X))$ and $\eta_g^3=\mathbb{E}\left[|g(X,X')|^3\right]$. By Callaert and Janssen (1978, Section 2), we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\sqrt{n}(L-\mathbb{E}[L])}{2\nu_{g_1}}\leqslant t\right\}-\Phi(t)\right|\leqslant C_{\mathsf{Ustat}}\cdot\frac{\eta_g^3}{\nu_{g_1}^3}\cdot\frac{1}{\sqrt{n}}\ , \tag{B.4}$$

for a universal constant C_{Ustat}. Next we bound the ratio $\frac{\eta_g^3}{\nu_{g_1}^3}$ in the following lemma, which is proved in Appendix D.6.

Lemma B.1. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Let $g(X, X')$ and $g_1(X)$ be defined as in the proof of Theorem 4.1. Then*

$$\nu_{g_1}^2 := \mathsf{Var}(g_1(X)) \geqslant \frac{1}{\pi^2} C_{\mathsf{variance}}^2$$

and

$$\nu_{g_1}^3 \leqslant \eta_g^3 := \mathbb{E}\left[|g(X, X')|^3\right] \leqslant C_{\mathsf{moment}}$$

where $C_{\mathsf{variance}}, C_{\mathsf{moment}}$ are constants depending only on $C_{\mathsf{cov}}, C_{\mathsf{kernel}}$ and not on (n, p_n, k_n).

In particular, this lemma implies that $S_{ab} := \pi\nu_{g_1}(\det(\Theta))^{-1} \geqslant C_{\mathsf{variance}} \cdot (\det(\Theta))^{-1}$.

Finally, the linear term L analysed here provides only an approximation to $\widetilde{\Theta}_{ab} - \Theta_{ab}$. Define

$$\Delta = \widetilde{\Theta}_{ab} - \Theta_{ab} - \frac{\pi}{2}(L - \mathbb{E}[L]) \ .$$

Then we have the bound

$$\begin{aligned} |\Delta| &= \left| u^\top \left[\frac{1}{2} \cdot \left(\frac{\pi}{2}\right)^2 \cdot \sin\left(\frac{\pi}{2}(t \cdot T + (1-t) \cdot \widehat{T})\right) \circ (\widehat{T} - T) \circ (\widehat{T} - T)\right] v \right| \\ &\leqslant ||u||_1 ||v||_1 ||\frac{1}{2} \cdot \left(\frac{\pi}{2}\right)^2 \cdot \sin\left(\frac{\pi}{2}(t \cdot T + (1-t) \cdot \widehat{T})\right) \circ (\widehat{T} - T) \circ (\widehat{T} - T)||_\infty \\ &\leqslant \frac{\pi^2}{8} ||u||_1 ||v||_1 ||\widehat{T} - T||_\infty^2 \\ &\leqslant \frac{\pi^2}{8} \cdot k_n \cdot (1 + 2C_{\mathsf{cov}} C_{\mathsf{sparse}})^2 \cdot ||\widehat{T} - T||_\infty^2 \ , \end{aligned} \tag{B.5}$$

where the last inequality holds by (B.2).

Finally, the next lemma is proved in de la Pena and Giné (1999).

Lemma B.2 ((de la Pena and Giné, 1999, Theorem 4.1.8))**.** *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$||\widehat{T} - T||_\infty \leqslant \sqrt{\frac{4 \log\left(2\binom{p_n}{2}/\delta\right)}{n}} \ .$$

Applying this lemma with $\delta = \frac{1}{2p_n}$, we have $||\widehat{T} - T||_\infty^2 \leqslant \frac{4\log(2p_n^3)}{n} \leqslant \frac{16\log(p_n)}{n}$ with probability at least $1 - \frac{1}{2p_n}$.

To summarize the computations so far, we have $\widetilde{\Theta}_{ab} - \Theta_{ab} = \frac{\pi}{2}(L - \mathbb{E}[L]) + \Delta$, where (B.4) gives an asymptotic normality result for the linear term $(L - \mathbb{E}[L])$, while (B.5) gives a bound on Δ. To prove therefore that $\widetilde{\Theta}_{ab} - \Theta_{ab}$ is asymptotically normal, we will use the following lemma (proved in Appendix D.1):

Lemma B.3. *Let A, B, C be random variables such that*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}\{A \leqslant t\} - \Phi(t)| \leqslant \epsilon_A \quad \text{and} \quad \mathbb{P}\{|B| \leqslant \delta_B, |C| \leqslant \delta_C\} \geqslant 1 - \epsilon_{BC} \ ,$$

where $\epsilon_A, \epsilon_{BC}, \delta_B, \delta_C \in (0, 1)$. Then the variable $(A + B) \cdot (1 + C)$ converges to a standard normal distribution with rate

$$\sup_{t \in \mathbb{R}} |\mathbb{P}\{(A + B) \cdot (1 + C) \leqslant t\} - \Phi(t)| \leqslant \delta_B + \frac{\delta_C}{1 - \delta_C} + \epsilon_A + \epsilon_{BC} \ .$$

We apply this lemma with $A = \frac{\pi}{2} \cdot \sqrt{n} \cdot \frac{L - \mathbb{E}[L]}{S_{ab} \cdot \det(\Theta)}$ and $B = \sqrt{n} \cdot \frac{\Delta}{S_{ab} \cdot \det(\Theta)}$ and $C = 0$. We have

$$\sup_{t \in \mathbb{R}} |\mathbb{P}\{A \leqslant t\} - \Phi(t)| \leqslant C_{\mathsf{Ustat}} \cdot \frac{C_{\mathsf{moment}}}{\left(\frac{1}{\pi^2} C^2_{\mathsf{variance}}\right)^{1.5}} \cdot \frac{1}{\sqrt{n}}$$

by (B.4) and Lemma B.1. Furthermore,

$$\mathbb{P}\left\{ |B| \leqslant \sqrt{n} \cdot \frac{\frac{\pi^2}{8} \cdot k_n \cdot (1 + 2C_{\mathsf{cov}} C_{\mathsf{sparse}})^2 \cdot \frac{16 \log(p_n)}{n}}{C_{\mathsf{variance}}} \right\} \leqslant \mathbb{P}\left\{ ||\widehat{T} - T||^2_\infty \leqslant \frac{16 \log(p_n)}{n} \right\} \geqslant 1 - \frac{1}{2p_n}$$

by (B.5) and Lemmas B.1 and B.2. Noting that $\sqrt{n} \cdot \frac{\widetilde{\Theta}_{ab} - \Theta_{ab}}{S_{ab}} = A + B$, and defining

$$C_{\mathsf{normal}} = \frac{2\pi^2 (1 + 2C_{\mathsf{cov}} C_{\mathsf{sparse}})^2}{C_{\mathsf{variance}}} + C_{\mathsf{Ustat}} \cdot \frac{C_{\mathsf{moment}}}{\left(\frac{1}{\pi^2} C^2_{\mathsf{variance}}\right)^{1.5}} ,$$

we have proved the desired result.

□

B.3 Proof of Theorem 4.2: gap between the estimator and the oracle estimator, and estimation of the variance

The first part of Theorem 4.2, which bounds the distance between our estimator $\breve{\Theta}$ of Θ and the oracle estimator $\widetilde{\Theta}$, is established using bounds on $\widehat{\Sigma} - \Sigma$ in Section 4.3. Details are given in Appendix B.3.1. The second part of Theorem 4.2, which bounds the error in estimating variance, $\left|\breve{S}_{ab} \cdot \det(\breve{\Theta}) - S_{ab} \cdot \det(\Theta)\right|$, is treated in Appendix B.3.2.

B.3.1 Bounds on $\breve{\Theta} - \widetilde{\Theta}$

Next we use our bounds on the covariance error, $\widehat{\Sigma} - \Sigma$, to derive a bound on the difference between our empirical estimator $\breve{\Theta}$ and the oracle estimator $\widetilde{\Theta}$ of Θ. The bounds we give here are deterministic. Write

$$\Delta_c = \breve{\gamma}_c - \gamma_c \text{ for } c = a, b .$$

Define the norm

$$||x||_{(k)} := \sqrt{||x||_2^2 + \frac{||x||_1^2}{k}} , \tag{B.6}$$

that is, $|| \cdot ||_{(k)}$ is the norm for which $\mathcal{B}_k$ is the unit ball.

The following lemma is proved in Appendix D.5:

Lemma B.4. *The following bound holds deterministically:*

$$||\breve{\Theta} - \widetilde{\Theta}||_\infty \leqslant \mathsf{C}(\Sigma) \cdot \max_{c \in \{a,b\}} ||\Delta_c||_2^2 +$$
$$\sup_{u,v \in \mathcal{B}_k} \left| u^\top (\widehat{\Sigma} - \Sigma) v \right| \cdot \left(2 \max_{c \in \{a,b\}} ||\Delta_c||_{(k)} \cdot \left(2 + \max_{c \in \{a,b\}} ||\gamma_c||_{(k)} \right) + \max_{c \in \{a,b\}} ||\Delta_c||^2_{(k)} \right) .$$

From this point on, we combine Assumptions 3.1, 3.2, and 3.3 with Corollary 4.8 and Lemma B.4 to obtain our probabilistic bound on $||\breve{\Theta} - \widetilde{\Theta}||_\infty$ (Theorem 4.2).

For $c = a, b$, applying (B.1) and (B.2),

$$||\gamma_c||_{(k_n)} \leqslant \sqrt{C_{\mathsf{cov}}^2 + \frac{(2C_{\mathsf{cov}}C_{\mathsf{sparse}}\sqrt{k_n})^2}{k_n}} = \sqrt{C_{\mathsf{cov}}^2 + 4C_{\mathsf{cov}}^2C_{\mathsf{sparse}}^2} \,. \tag{B.7}$$

Next, for $c = a, b$, by Assumption 3.3, with probability at least $1 - \delta_n$,

$$\begin{aligned} ||\Delta_c||_{(k_n)} &= \sqrt{||\Delta_c||_2^2 + \frac{||\Delta_c||_1^2}{k_n}} \\ &\leqslant \sqrt{\left(C_{\mathsf{est}}\sqrt{\frac{k_n\log(p_n)}{n}}\right)^2 + \frac{\left(C_{\mathsf{est}}\sqrt{\frac{k_n^2\log(p_n)}{n}}\right)^2}{k_n}} = C_{\mathsf{est}}\sqrt{\frac{2k_n\log(p_n)}{n}} \,. \end{aligned} \tag{B.8}$$

Next, we use Corollary 4.8. Setting $\delta_1 = \delta_2 = \frac{1}{6p_n}$, we see that by the assumption $p_n \geqslant 2, k_n \geqslant 1$ and the assumption $n \geqslant 15k_n\log(p_n)$ stated in Theorem 4.2, the conditions of Corollary 4.8 must hold. Then, with probability at least $1 - \delta_1 - \delta_2 = 1 - \frac{1}{3p_n}$,

$$\begin{aligned} &\sup_{u,v\in\mathcal{B}_{k_n}} \left|u^\top(\widehat{\Sigma} - \Sigma)v\right| \\ &\leqslant \frac{\pi^2}{8}\cdot k_n \cdot \frac{4\log\left(12p_n\binom{p_n}{2}\right)}{n} + 2\pi\cdot 16(1+\sqrt{5})C_{\mathsf{cov}}\cdot\sqrt{\frac{\log(12p_n) + (k_n+1)\log(12p_n)}{n}} \\ &\leqslant C_{\mathsf{cov}}\cdot C'\cdot\sqrt{\frac{k_n\log(p_n)}{n}} \,, \end{aligned} \tag{B.9}$$

where we choose the universal constant $C' = 3\pi^2 + 2\pi\cdot 16(1+\sqrt{5})\sqrt{15}$ which guarantees that the last inequality holds (using the assumptions $n \geqslant k_n\log(p_n)$ and $p_n \geqslant 2$).

Finally, applying the deterministic bound in Lemma B.4, we see that with probability at least $1 - \frac{1}{3p_n}$, on the event that the bounds (3.1) in Assumption 3.3 hold,

$$\begin{aligned} ||\breve{\Theta} - \widetilde{\Theta}||_\infty \leqslant C_{\mathsf{cov}}\cdot\max_{c\in\{a,b\}}||\Delta_c||_2^2 + \sup_{u,v\in\mathcal{B}_{k_n}}\left|u^\top(\widehat{\Sigma} - \Sigma)v\right|\cdot \\ \left(2\max_{c\in\{a,b\}}||\Delta_c||_{(k_n)}\cdot\left(2 + \max_{c\in\{a,b\}}||\gamma_c||_{(k_n)}\right) + \max_{c\in\{a,b\}}||\Delta_c||_{(k_n)}^2\right). \end{aligned}$$

Applying Assumption 3.2 and calculations (B.7), (B.8), and (B.9) above,

$$\begin{aligned} &\leqslant C_{\mathsf{cov}}\cdot C_{\mathsf{est}}^2\frac{k_n\log(p_n)}{n} + C_{\mathsf{cov}}\cdot C'\cdot\sqrt{\frac{k_n\log(p_n)}{n}}\cdot \\ &\qquad \left(2C_{\mathsf{est}}\sqrt{\frac{2k_n\log(p_n)}{n}}\cdot\left(2 + \sqrt{C_{\mathsf{cov}}^2 + 4C_{\mathsf{cov}}^2C_{\mathsf{sparse}}^2}\right) + C_{\mathsf{est}}^2\frac{2k_n\log(p_n)}{n}\right) \\ &\leqslant \frac{k_n\log(p_n)}{n}\cdot\left[C_{\mathsf{cov}}C_{\mathsf{est}}^2 + C_{\mathsf{cov}}\cdot C'\cdot\left(2\sqrt{2}\cdot C_{\mathsf{est}}\cdot\left(2 + \sqrt{C_{\mathsf{cov}}^2 + 4C_{\mathsf{cov}}^2C_{\mathsf{sparse}}^2}\right) + 2C_{\mathsf{est}}^2\right)\right] , \end{aligned}$$

where the last step uses the fact that $\frac{k_n \log(p_n)}{n} \leqslant \sqrt{\frac{k_n \log(p_n)}{n}}$ (which is true by assumption in Theorem 4.2). Defining C_{oracle} to be at least as large as the expression in square brackets above, we have completed the proof of the first part of Theorem 4.2.

B.3.2 Variance estimate

For the second part of the theorem, that is, bounding the error in the variance estimate $\check{S}_{ab}$, we state this bound as a lemma and defer the proof to Appendix D.9, since we need to develop some additional technical results before treating this bound.

Lemma B.5. *Under the assumptions and definitions of Theorem 4.2, with probability at least* $1 - \frac{1}{6p_n}$*, if* $n \geqslant k_n^2 \log(p_n)$*, on the event that the bounds* (3.1) *in Assumption 3.3 hold,*

$$\left|\check{S}_{ab} \cdot \det(\check{\Theta}) - S_{ab} \cdot \det(\Theta)\right| \leqslant C_{\mathsf{oracle}} \cdot \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ .$$

Combining this lemma with the work above, and using Assumption 3.3, we have proved that both bounds stated in Theorem 4.2 hold with probability at least $1 - \frac{1}{p_n} - \delta_n$, as desired.

B.4 Proof of Theorem 3.5: main result

We now prove our main result, Theorem 3.5.

Proof of Theorem 3.5. Recall that our goal is to prove that $\frac{\sqrt{n}(\check{\Omega}_{ab} - \Omega_{ab})}{\check{S}_{ab}}$ converges to the $N(0, 1)$ distribution. Recalling that $\Theta = (\Omega_{ab,ab})^{-1}$ and using the formula for a 2×2 matrix inverse, we separate this random variable into several terms:

$$\begin{aligned}
\frac{\sqrt{n}(\check{\Omega}_{ab} - \Omega_{ab})}{\check{S}_{ab}} &= \frac{\sqrt{n}\left(\frac{-\check{\Theta}_{ab}}{\det(\check{\Theta})} - \frac{-\Theta_{ab}}{\det(\Theta)}\right)}{\check{S}_{ab}} = \frac{\sqrt{n}\left(-\check{\Theta}_{ab} + \Theta_{ab} \cdot \frac{\det(\check{\Theta})}{\det(\Theta)}\right)}{\check{S}_{ab} \cdot \det(\check{\Theta})} \\
&= \frac{\sqrt{n}\left(\Theta_{ab} - \widetilde{\Theta}_{ab} + \widetilde{\Theta}_{ab} - \check{\Theta}_{ab} - \Theta_{ab} \cdot \left(1 - \frac{\det(\check{\Theta})}{\det(\Theta)}\right)\right)}{\check{S}_{ab} \cdot \det(\check{\Theta})} \\
&= \left[-\frac{\sqrt{n}\left(\widetilde{\Theta}_{ab} - \Theta_{ab}\right)}{S_{ab} \cdot \det(\Theta)} + \frac{\sqrt{n}\left(\widetilde{\Theta}_{ab} - \check{\Theta}_{ab}\right)}{S_{ab} \cdot \det(\Theta)} + \frac{\sqrt{n} \cdot \Omega_{ab} \cdot \left(\det(\Theta) - \det(\check{\Theta})\right)}{S_{ab} \cdot \det(\Theta)}\right] \\
&\qquad \times \left[1 + \frac{\check{S}_{ab} \cdot \det(\check{\Theta}) - S_{ab} \cdot \det(\Theta)}{S_{ab} \cdot \det(\Theta)}\right] \ .
\end{aligned}$$

To show that $\frac{\sqrt{n}(\check{\Omega}_{ab} - \Omega_{ab})}{\check{S}_{ab}}$ converges to the standard normal distribution, we will can apply Lemma B.3 (stated in Appendix B.2). In order to apply this lemma and obtain the desired result, we assemble the following pieces:

First, the variable $A := -\frac{\sqrt{n}(\widetilde{\Theta}_{ab} - \Theta_{ab})}{S_{ab} \cdot \det(\Theta)}$ satisfies $\sup_{t \in \mathbb{R}} |\mathbb{P}\{A \leqslant t\} - \Phi(t)| \leqslant C_{\mathsf{normal}} \cdot \frac{k_n \log(p_n)}{\sqrt{n}} + \frac{1}{2p_n}$, as shown in Theorem 4.1.

Second, we define variables $B := \frac{\sqrt{n}\left(\widetilde{\Theta}_{ab}-\breve{\Theta}_{ab}\right)}{S_{ab}\cdot\det(\Theta)} + \frac{\sqrt{n}\cdot\Omega_{ab}\cdot\left(\det(\Theta)-\det(\breve{\Theta})\right)}{S_{ab}\cdot\det(\Theta)}$ and $C := \frac{\breve{S}_{ab}\cdot\det(\breve{\Theta})-S_{ab}\cdot\det(\Theta)}{S_{ab}\cdot\det(\Theta)}$, and set

$$\delta_B = \frac{k_n\log(p_n)}{\sqrt{n}} \cdot \left(\frac{C_{\mathsf{oracle}} + 4C_{\mathsf{cov}}^2 C_{\mathsf{oracle}} + 2C_{\mathsf{cov}}C_{\mathsf{oracle}}^2}{C_{\mathsf{variance}}}\right)$$

and

$$\delta_C = \frac{C_{\mathsf{oracle}}}{C_{\mathsf{variance}}} \cdot \sqrt{\frac{k_n^2\log(p_n)}{n}} \; .$$

We now show that, by Theorem 4.2, with probability at least $1-\frac{1}{2p_n}-\delta_n$ it holds that $|B| \leqslant \delta_B$ and $|C| \leqslant \delta_C$. For the variable C, this is a trivial consequence of the bound on $\left|\breve{S}_{ab}\cdot\det(\breve{\Theta}) - S_{ab}\cdot\det(\Theta)\right|$ in Theorem 4.2 combined with the lower bound $S_{ab}\cdot\det(\Theta) \geqslant C_{\mathsf{variance}}$ from Theorem 4.1.

Now we turn to the bound on B. To prove this bound, observe that $||\breve{\Theta}-\widetilde{\Theta}||_\infty \leqslant C_{\mathsf{oracle}}\cdot\frac{k_n\log(p_n)}{n}$ by Theorem 4.2 (with the stated probability). We also have

$$\left|\frac{\sqrt{n}\left(\widetilde{\Theta}_{ab}-\breve{\Theta}_{ab}\right)}{S_{ab}\cdot\det(\Theta)}\right| \leqslant \sqrt{n}\cdot\frac{1}{S_{ab}\cdot\det(\Theta)}\cdot||\breve{\Theta}-\widetilde{\Theta}||_\infty \leqslant \frac{\sqrt{n}}{C_{\mathsf{variance}}}\cdot||\breve{\Theta}-\widetilde{\Theta}||_\infty \; ,$$

where the last step follows from Theorem 4.1. And,

$$\begin{aligned}\left|\det(\breve{\Theta})-\det(\Theta)\right| &= \left|\left(\breve{\Theta}_{aa}\breve{\Theta}_{bb}-\breve{\Theta}_{ab}^2\right)-\left(\Theta_{aa}\Theta_{bb}-\Theta_{ab}^2\right)\right| \\ &\leqslant 4||\Theta||_\infty||\breve{\Theta}-\Theta||_\infty + 2||\breve{\Theta}-\Theta||_\infty^2\end{aligned}$$

and

$$|\Omega_{ab}| \leqslant \lambda_{\max}(\Omega) = (\lambda_{\min}(\Sigma))^{-1} \leqslant C_{\mathsf{cov}} \; .$$

Therefore,

$$\begin{aligned}\left|\frac{\sqrt{n}\cdot\Omega_{ab}\cdot\left(\det(\Theta)-\det(\breve{\Theta})\right)}{S_{ab}\cdot\det(\Theta)}\right| &\leqslant \sqrt{n}\cdot\frac{|\Omega_{ab}|}{S_{ab}\cdot\det(\Theta)}\cdot\left(4||\Theta||_\infty||\breve{\Theta}-\Theta||_\infty + 2||\breve{\Theta}-\Theta||_\infty^2\right) \\ &\leqslant \sqrt{n}\cdot\frac{C_{\mathsf{cov}}}{C_{\mathsf{variance}}}\cdot\left(4\mathsf{C}_{\mathsf{cov}}||\breve{\Theta}-\Theta||_\infty + 2||\breve{\Theta}-\Theta||_\infty^2\right) \; ,\end{aligned}$$

where the last step follows from Theorem 4.1 along with the fact that

$$||\Theta||_\infty \leqslant \lambda_{\max}(\Theta) = (\lambda_{\min}(\Omega_{ab,ab}))^{-1} \leqslant (\lambda_{\min}(\Omega))^{-1} = \lambda_{\max}(\Sigma) \leqslant \mathsf{C}_{\mathsf{cov}} \; .$$

Combining everything, we have

$$\begin{aligned}|B| &\leqslant \sqrt{n}\cdot\frac{1}{C_{\mathsf{variance}}}\cdot||\breve{\Theta}-\widetilde{\Theta}||_\infty + \sqrt{n}\cdot\frac{C_{\mathsf{cov}}}{C_{\mathsf{variance}}}\cdot\left(4\mathsf{C}_{\mathsf{cov}}||\breve{\Theta}-\Theta||_\infty + 2||\breve{\Theta}-\Theta||_\infty^2\right) \\ &\leqslant \frac{k_n\log(p_n)}{\sqrt{n}}\left[\frac{C_{\mathsf{oracle}} + 4C_{\mathsf{cov}}^2C_{\mathsf{oracle}} + 2C_{\mathsf{cov}}C_{\mathsf{oracle}}^2\cdot\frac{k_n\log(p_n)}{n}}{C_{\mathsf{variance}}}\right] \; .\end{aligned}$$

If $n < k_n\log(p_n)$, then the main result in Theorem 3.5 holds trivially. Assuming then that $n \geqslant k_n\log(p_n)$, we have proved the desired bound on $|B|$.

Given these convergence results, we apply Lemma B.3 to obtain the following result:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\sqrt{n}(\check{\Omega}_{ab}-\Omega_{ab})}{\check{S}_{ab}}\leqslant t\right\}-\Phi(t)\right|\leqslant\delta_B+\frac{\delta_C}{1-\delta_C}+\epsilon_A+\epsilon_{BC}$$

$$=\frac{k_n\log(p_n)}{\sqrt{n}}\cdot\left(\frac{C_{\mathsf{oracle}}+4C_{\mathsf{cov}}^2C_{\mathsf{oracle}}+2C_{\mathsf{cov}}C_{\mathsf{oracle}}^2}{C_{\mathsf{variance}}}\right)+$$

$$\frac{\frac{C_{\mathsf{oracle}}}{C_{\mathsf{variance}}}\cdot\sqrt{\frac{k_n^2\log(p_n)}{n}}}{1-\frac{C_{\mathsf{oracle}}}{C_{\mathsf{variance}}}\cdot\sqrt{\frac{k_n^2\log(p_n)}{n}}}+C_{\mathsf{normal}}\cdot\frac{k_n\log(p_n)}{\sqrt{n}}+\frac{1}{2p_n}+\frac{1}{2p_n}+\delta_n\ .$$

If $\frac{C_{\mathsf{oracle}}}{C_{\mathsf{variance}}}\cdot\sqrt{\frac{k_n^2\log(p_n)}{n}}>\frac{1}{2}$,, then the result of Theorem 3.5 holds trivially, and so assuming that this is not the case, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\sqrt{n}(\check{\Omega}_{ab}-\Omega_{ab})}{\check{S}_{ab}}\leqslant t\right\}-\Phi(t)\right|\leqslant C_{\mathsf{converge}}\cdot\sqrt{\frac{k_n^2\log^2(p_n)}{n}}+\frac{1}{p_n}+\delta_n\ ,$$

where

$$C_{\mathsf{converge}}:=\frac{C_{\mathsf{oracle}}+4C_{\mathsf{cov}}^2C_{\mathsf{oracle}}+2C_{\mathsf{cov}}C_{\mathsf{oracle}}^2}{C_{\mathsf{variance}}}+\frac{2C_{\mathsf{oracle}}}{C_{\mathsf{variance}}}+C_{\mathsf{normal}}\ .$$

□

C Accuracy of the initial Lasso estimator

Proof of Corollary 3.8. Define

$$A=\widehat{\Sigma}_I,\ z=\widehat{\Sigma}_{Ia},\ x^\star=\gamma_a,\ p=p_n-2,\ k=k_n\ .$$

Now we apply Theorem 3.7 to this sparse recovery problem. In order to do so, we need to check that the conditions (3.3), (3.4), and (3.5) hold, and that γ_a is feasible under the condition $||\gamma||_1\leqslant R$. Once these conditions are satisfied, the result of Theorem 3.7 can be applied to this setting.

Feasibility of γ_a. Define $R=C_{\mathsf{cov}}\sqrt{2k_n}$. As proved in (B.1), $||\gamma_a||_2\leqslant C_{\mathsf{cov}}$, and furthermore $||\gamma_a||_0\leqslant||\Omega_a||_0+||\Omega_b||_0\leqslant 2k_n$ (this is true because $\gamma_a=-\Omega_{I,ab}\Theta_{ab,a}$ by (B.2)). Therefore, $||\gamma_a||_1\leqslant C_{\mathsf{cov}}\sqrt{2k_n}=R$.

Condition (3.4) (restricted strong convexity). Now we need to check that the restricted strong convexity conditions (3.3) hold for our matrix $A=\widehat{\Sigma}_I$. By Corollary 4.8, there exists a constant C_{RSC} depending only on C_{cov}, such that if $n\geqslant 16\log(p_n)$, then with probability at least $1-\frac{1}{8p_n}$, for all $v\in\mathbb{R}^{p_n}$,

$$\left|v^\top\left(\widehat{\Sigma}_I-\Sigma_I\right)v\right|\leqslant\frac{1}{2C_{\mathsf{cov}}}\left(||v||_2^2+||v||_1^2\cdot\frac{C_{\mathsf{RSC}}\log(p_n)}{n}\right)\ .$$

(To see this, apply Corollary 4.8 with $\frac{n}{C_{\mathsf{RSC}}\log(p_n)}$ in place of k_n; the assumption $n \geqslant 16\log(p_n)$ ensures that we can choose C_{RSC} so that the condition of Corollary 4.8 is satisfied.) Then, if this event holds, for all $v \in \mathbb{R}^I$ we have

$$v^\top \widehat{\Sigma}_I v \geqslant v^\top \Sigma_I v - \left| v^\top \left(\widehat{\Sigma}_I - \Sigma_I\right) v \right| \geqslant C_{\mathsf{cov}}^{-1} \cdot ||v||_2^2 - \left| v^\top \left(\widehat{\Sigma}_I - \Sigma_I\right) v \right|$$
$$\geqslant \frac{1}{2C_{\mathsf{cov}}} \cdot ||v||_2^2 - \frac{C_{\mathsf{RSC}}}{C_{\mathsf{cov}}} \cdot \frac{\log(p_n)}{n} \cdot ||v||_1^2 \,.$$

Therefore, with probability at least $1 - \frac{1}{8p_n}$, the restricted strong convexity condition (3.3) holds with

$$\alpha_1 = \frac{1}{2C_{\mathsf{cov}}} \text{ and } \tau_1 = \frac{C_{\mathsf{RSC}}}{C_{\mathsf{cov}}} \,.$$

Condition (3.5) (penalty parameter). Below, we will prove that, with probability at least $1 - \frac{3}{8p_n}$,

$$||Ax^\star - z||_\infty = ||\widehat{\Sigma}_I \gamma_a - \widehat{\Sigma}_{Ia}||_\infty \leqslant \frac{\pi}{2} C_{\mathsf{feasible}} \sqrt{\frac{\log(p_n)}{n}} + \sqrt{\frac{\log(p_n)}{n}} \cdot \left[\frac{1.5\sqrt{3}\pi^2\sqrt{1 + C_{\mathsf{cov}}^2}}{\sqrt{C_{\mathsf{sample}}}}\right] \,, \quad \text{(C.1)}$$

for a constant C_{feasible} depending only on C_{cov}, as long as we set

$$C_{\mathsf{sample}} \geqslant \left[16(1+\sqrt{5})C_{\mathsf{cov}}\sqrt{1 + C_{\mathsf{cov}}^2}\right]^2 \,.$$

Given that this is true, we now require that condition (3.5) holds, that is,

$$\max\left\{4||Ax^\star - z||_\infty, 4\alpha_1\sqrt{\frac{\log(p)}{n}}\right\} \leqslant \lambda \leqslant \frac{\alpha_1}{6R} \,.$$

Define

$$C_{\mathsf{Lasso}} = \max\left\{4\left[\frac{\pi}{2}C_{\mathsf{feasible}} + \frac{1.5\sqrt{3}\pi^2\sqrt{1 + C_{\mathsf{cov}}^2}}{\sqrt{C_{\mathsf{sample}}}}\right], \frac{2}{C_{\mathsf{cov}}}\right\} \,,$$

Plugging in the bound (C.1), we see that the lower bound on λ is satisfied for $\lambda = C_{\mathsf{Lasso}}\sqrt{\frac{\log(p_n)}{n}}$. To check the upper bound, we only need

$$\lambda = C_{\mathsf{Lasso}}\sqrt{\frac{\log(p_n)}{n}} \leqslant \frac{\alpha_1}{6R} = \frac{1}{2C_{\mathsf{cov}}^2\sqrt{2k_n}}.$$

Assuming that

$$n \geqslant 8C_{\mathsf{Lasso}}^2 C_{\mathsf{cov}}^4 \cdot k_n \log(p_n) \,, \quad \text{(C.2)}$$

then this follows directly. Therefore, (3.5) is satisfied with probability at least $1 - \frac{3}{8p_n}$.

Condition (3.4) (sample size). To satisfy (3.4), by plugging in the definitions of R, α_1, and τ_1 above, we see that it is sufficient to require

$$n \geqslant 64C_{\mathsf{cov}}^2 C_{\mathsf{RSC}} \max\{1, 2C_{\mathsf{RSC}}\} \cdot k_n \log(p_n) \,. \quad \text{(C.3)}$$

Conclusion. Combining all of our work above, we see that the conditions (3.3), (3.4), and (3.5), and the feasibility of γ_a, are all satisfied with probability at least $1 - \frac{1}{2p_n}$, as long as

$$n \geqslant C_{\textsf{sample}} k_n \log(p_n)$$

for

$$C_{\textsf{sample}} := \max\left\{16, \left[16(1+\sqrt{5})C_{\textsf{cov}}\sqrt{1+C_{\textsf{cov}}^2}\right]^2, 8C_{\textsf{Lasso}}^2 C_{\textsf{cov}}^4, 64C_{\textsf{cov}}^2 C_{\textsf{RSC}} \max\{1, 2C_{\textsf{RSC}}\}\right\} .$$

Therefore, applying Theorem 3.7, if these high probability events hold, then then for any $\breve{\gamma}_a$ that is a local minimizer of

$$L(x) = \frac{1}{2}\gamma^\top \widehat{\Sigma}_I \gamma - \gamma^\top \widehat{\Sigma}_{Ia} + \lambda||\gamma||_1$$

over the set $\{\gamma \in \mathbb{R}^I : ||\gamma||_1 \leqslant 2C_{\textsf{cov}}C_{\textsf{sparse}}\sqrt{k_n}\}$, it holds that

$$||\breve{\gamma}_a - \gamma_a||_2 \leqslant \frac{1.5\lambda \cdot \sqrt{2k_n}}{\alpha_1} = 3\sqrt{2}C_{\textsf{cov}}\lambda\sqrt{k_n} \text{ and } ||\breve{\gamma}_a - \gamma_a||_1 \leqslant \frac{6\lambda \cdot 2k_n}{\alpha_1} = 24C_{\textsf{cov}}\lambda\sqrt{k_n} .$$

By the same arguments, the same results hold for estimating γ_b.

Proving (C.1) Now we consider the term $||Ax^\star - z||_\infty = ||\widehat{\Sigma}_I\gamma_a - \widehat{\Sigma}_{Ia}||_\infty$. Since $\gamma_a = \Sigma_I^{-1}\Sigma_{Ia}$, we have

$$||\widehat{\Sigma}_I\gamma_a - \widehat{\Sigma}_{Ia}||_\infty = ||(\widehat{\Sigma}_I - \Sigma_I)\gamma_a - (\widehat{\Sigma}_{Ia} - \Sigma_{Ia})||_\infty = ||(\widehat{\Sigma} - \Sigma)u||_\infty ,$$

where $u \in \mathbb{R}^{p_n}$ is the fixed vector with

$$u_a = 1, u_b = 0, u_I = -\gamma_a .$$

By the Taylor expansion of $\widehat{\Sigma} - \Sigma$ (calculated as in (B.3)), we have

$$\begin{aligned}
||\widehat{\Sigma}_I\gamma_a - \widehat{\Sigma}_{Ia}||_\infty = ||(\widehat{\Sigma} - \Sigma)u||_\infty &= \max_j \left|\mathbf{e}_j^\top(\widehat{\Sigma} - \Sigma)u\right| \\
&\leqslant \frac{\pi}{2}\max_j\left|\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T} - T)\right)u\right| + \left|\frac{\pi^2}{8}\mathbf{e}_j^\top\left(\sin\left(\frac{\pi}{2}\overline{T}\right)\circ(\widehat{T} - T)\circ(\widehat{T} - T)\right)u\right| \\
&\leqslant \frac{\pi}{2}\max_j\left|\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T} - T)\right)u\right| + \frac{\pi^2}{8}||u||_1||\widehat{T} - T||_\infty^2 . \qquad \text{(C.4)}
\end{aligned}$$

Next we bound each term in this final expression (C.4) separately. Beginning with the second term, by (B.1), we know that $||u||_1 \leqslant \sqrt{||u||_0}||u||_2 \leqslant \sqrt{1+2k_n} \cdot \sqrt{1+C_{\textsf{cov}}^2} \leqslant \sqrt{k_n} \cdot \sqrt{3(1+C_{\textsf{cov}}^2)}$, where to bound $||u||_0$ we use the calculation $||\gamma_a||_0 \leqslant 2k_n$ from before. Furthermore, by Lemma B.2, with probability at least $1 - \frac{1}{8p_n}$,

$$||\widehat{T} - T||_\infty \leqslant \sqrt{\frac{12\log(8p_n)}{n}} \leqslant \sqrt{\frac{48\log(p_n)}{n}} ,$$

using $p_n \geqslant 2$. Therefore, the second term in (C.4) is bounded as

$$\frac{\pi^2}{8}||u||_1||\widehat{T} - T||_\infty^2 \leqslant \frac{\pi^2}{8}\sqrt{k_n} \cdot \sqrt{3(1+C_{\textsf{cov}}^2)}\frac{48\log(p_n)}{n} \leqslant \sqrt{\frac{\log(p_n)}{n}} \cdot \left[\frac{6\sqrt{3}\pi^2\sqrt{1+C_{\textsf{cov}}^2}}{\sqrt{C_{\textsf{sample}}}}\right] , \qquad \text{(C.5)}$$

where we use the assumption $n \geqslant C_{\mathsf{sample}} \log(p_n)$.

Next we turn to the first term in (C.4). In order to bound this term, we begin by stating two lemmas (proved in Appendix D.7):

Lemma C.1. *There exist vectors $a_1, a_2, \dots$ and $b_1, b_2, \dots$ with $||a_r||_\infty, ||b_r||_\infty \leqslant 1$ for all $r \geqslant 1$, and a sequence $t_1, t_2, \dots \geqslant 0$ with $\sum_r t_r = 4$, such that $\cos\left(\frac{\pi}{2}T\right) = \sum_{r\geqslant 1} t_r a_r b_r^\top$.*

Lemma C.2. *For fixed u, v with $||u||_2, ||v||_2 \leqslant 1$, for any $|t| \leqslant \frac{n}{4(1+\sqrt{5})C_{\mathsf{cov}}}$,*

$$\mathbb{E}\left[\exp\left(t \cdot u^\top(\widehat{T} - T)v\right)\right] \leqslant \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 t^2 \cdot C_{\mathsf{cov}}^2}{n}\right) .$$

By Lemma C.1, we can write

$$\cos\left(\frac{\pi}{2}T\right) = \sum_r t_r \cdot a_r b_r^\top ,$$

where $t_r \geqslant 0$, $\sum_r t_r = 4$, and $||a_r||_\infty, ||b_r||_\infty \leqslant 1$. Then

$$\mathbf{e}_j^\top \left(\cos\left(\frac{\pi}{2}T\right) \circ (\widehat{T} - T)\right) u = \left\langle \cos\left(\frac{\pi}{2}T\right) \circ \mathbf{e}_j u^\top, \widehat{T} - T \right\rangle = \sum_r t_r \cdot (a_r \circ \mathbf{e}_j)^\top (\widehat{T} - T)(b_r \circ u) .$$

Note that

$$||a_r \circ \mathbf{e}_j||_2 \leqslant ||a_r||_\infty \cdot ||\mathbf{e}_j||_2 \leqslant 1$$

and, by (B.1),

$$||b_r \circ u||_2 \leqslant ||b_r||_\infty \cdot ||u||_2 \leqslant \sqrt{1 + C_{\mathsf{cov}}^2} .$$

Then for any $|t| \leqslant \frac{n}{16(1+\sqrt{5})C_{\mathsf{cov}}\sqrt{1+C_{\mathsf{cov}}^2}}$,

$$\mathbb{E}\left[\exp\left\{t \cdot \mathbf{e}_j^\top \left(\cos\left(\frac{\pi}{2}T\right) \circ (\widehat{T} - T)\right) u\right\}\right] = \mathbb{E}\left[\exp\left\{\sum_r t_r \left[t \cdot (a_r \circ \mathbf{e}_j)^\top (\widehat{T} - T)(b_r \circ u)\right]\right\}\right] .$$

By convexity of the function $x \mapsto e^x$,

$$= \sum_r \frac{t_r}{4} \mathbb{E}\left[\exp\left\{4\left[t \cdot (a_r \circ \mathbf{e}_j)^\top (\widehat{T} - T)(b_r \circ u)\right]\right\}\right] .$$

By Lemma C.2,

$$\leqslant \sum_r \frac{t_r}{4} \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 16 t^2 \cdot C_{\mathsf{cov}}^2 (1 + C_{\mathsf{cov}}^2)}{n}\right)$$

$$= \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 16 t^2 \cdot C_{\mathsf{cov}}^2 (1 + C_{\mathsf{cov}}^2)}{n}\right) .$$

Observe that we can set $t = \pm\sqrt{n\log(p_n)}$, which satisfies $|t| \leqslant \frac{n}{16(1+\sqrt{5})C_{\mathsf{cov}}\sqrt{1+C_{\mathsf{cov}}^2}}$ as long as we set $C_{\mathsf{sample}} \geqslant \left[16(1+\sqrt{5})C_{\mathsf{cov}}\sqrt{1+C_{\mathsf{cov}}^2}\right]^2$, due to the assumption $n \geqslant C_{\mathsf{sample}}\log(p_n)$. Then, we see that for any $C > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)\right)u\right| > C\sqrt{\frac{\log(p_n)}{n}}\right\}\\
&\qquad\qquad\leqslant \mathbb{E}\left[e^{\sqrt{n\log(p_n)}\cdot\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)\right)u - \sqrt{n\log(p_n)}\cdot C\sqrt{\frac{\log(p_n)}{n}}}\right]\\
&\qquad\qquad+\mathbb{E}\left[e^{-\sqrt{n\log(p_n)}\cdot\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)\right)u - \sqrt{n\log(p_n)}\cdot C\sqrt{\frac{\log(p_n)}{n}}}\right]\\
&\leqslant 2\exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 16(\sqrt{n\log(p_n)})^2\cdot C_{\mathsf{cov}}^2(1+C_{\mathsf{cov}}^2)}{n}\right)\cdot\exp\left\{-\sqrt{n\log(p_n)}\cdot C\sqrt{\frac{\log(p_n)}{n}}\right\}\\
&\qquad\qquad\qquad\qquad\leqslant 2p_n^{-\left(C-\left[4(1+\sqrt{5})\right]^2\cdot 16C_{\mathsf{cov}}^2(1+C_{\mathsf{cov}}^2)\right)} = 2p_n^{-5} \leqslant \frac{1}{4p_n^2}\ ,
\end{aligned}$$

where we set $C = C_{\mathsf{feasible}} := 5 + \left[4(1+\sqrt{5})\right]^2\cdot 16C_{\mathsf{cov}}^2(1+C_{\mathsf{cov}}^2)$. Therefore,

$$\mathbb{P}\left\{\max_j\left|\mathbf{e}_j^\top\left(\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}-T)\right)u\right| > C_{\mathsf{feasible}}\sqrt{\frac{\log(p_n)}{n}}\right\} \leqslant \frac{1}{4p_n}\ . \tag{C.6}$$

Combining (C.5) and (C.6), and returning to (C.4), we have

$$||\widehat{\Sigma}_I\gamma_a - \widehat{\Sigma}_{Ia}||_\infty \leqslant \frac{\pi}{2}C_{\mathsf{feasible}}\sqrt{\frac{\log(p_n)}{n}} + \sqrt{\frac{\log(p_n)}{n}}\cdot\left[\frac{1.5\sqrt{3}\pi^2\sqrt{1+C_{\mathsf{cov}}^2}}{\sqrt{C_{\mathsf{sample}}}}\right]\ ,$$

with probability at least $1 - \frac{3}{8p_n}$. This proves (C.1). □

D Proofs of lemmas

D.1 Proof of the normal convergence lemma

Lemma B.3. *Let A, B, C be random variables such that*

$$\sup_{t\in\mathbb{R}}|\mathbb{P}\{A \leqslant t\} - \Phi(t)| \leqslant \epsilon_A \quad \text{and} \quad \mathbb{P}\{|B| \leqslant \delta_B, |C| \leqslant \delta_C\} \geqslant 1 - \epsilon_{BC}\ ,$$

where $\epsilon_A, \epsilon_{BC}, \delta_B, \delta_C \in [0,1)$. Then the variable $(A+B)\cdot(1+C)$ converges to a standard normal distribution with rate

$$\sup_{t\in\mathbb{R}}|\mathbb{P}\{(A+B)\cdot(1+C) \leqslant t\} - \Phi(t)| \leqslant \delta_B + \frac{\delta_C}{1-\delta_C} + \epsilon_A + \epsilon_{BC}\ .$$

Proof of Lemma B.3. First, define truncated versions of B and C:

$$\widetilde{B} = \mathrm{sign}(B)\cdot\min\{|B|, \delta_B\},\ \widetilde{C} = \mathrm{sign}(C)\cdot\min\{|C|, \delta_C\}\ .$$

Then, for any $t \in \mathbb{R}$,

$$\left|\mathbb{P}\left\{(A+B)\cdot(1+C) \leqslant t\right\} - \mathbb{P}\left\{(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant t\right\}\right| \leqslant \mathbb{P}\left\{B \neq \widetilde{B} \text{ or } C \neq \widetilde{C}\right\} \leqslant \epsilon_{BC} \ .$$

Note that $|\widetilde{B}| \leqslant \delta_B$ and $|\widetilde{C}| \leqslant \delta_C$ with probability 1.

Next, fix any $t \geqslant 0$ and suppose that $A \leqslant \frac{t}{1+\delta_C} - \delta_B$. Then

$$(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant \left(\left(\frac{t}{1+\delta_C} - \delta_B\right) + \delta_B\right)\cdot(1+\delta_C) = t \ ,$$

and so

$$\begin{aligned}
&\mathbb{P}\left\{(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant t\right\} \geqslant \mathbb{P}\left\{A \leqslant \frac{t}{1+\delta_C} - \delta_B\right\} \\
&\geqslant \Phi\left(\frac{t}{1+\delta_C} - \delta_B\right) - \epsilon_A \\
&= \Phi(t) - \mathbb{P}\left\{\frac{t}{1+\delta_C} - \delta_B < N(0,1) < \frac{t}{1+\delta_C}\right\} - \mathbb{P}\left\{\frac{t}{1+\delta_C} < N(0,1) < t\right\} - \epsilon_A \ .
\end{aligned}$$

Since the density of the normal distribution is bounded by $\frac{1}{\sqrt{2\pi}} \leqslant 1$,

$$\geqslant \Phi(t) - \delta_B - \mathbb{P}\left\{\frac{t}{1+\delta_C} < N(0,1) < t\right\} - \epsilon_A.$$

Applying Lemma D.1 (stated below),

$$\geqslant \Phi(t) - \delta_B - \delta_C - \epsilon_A \ .$$

To prove the reverse bound, suppose that $(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant t$. Then

$$A = \frac{(A+\widetilde{B})\cdot(1+\widetilde{C})}{1+\widetilde{C}} - \widetilde{B} \leqslant \frac{t}{1-\delta_C} + \delta_B \ .$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant t\right\} \leqslant \mathbb{P}\left\{A \leqslant \frac{t}{1-\delta_C} + \delta_B\right\} \\
&\leqslant \Phi\left(\frac{t}{1-\delta_C} + \delta_B\right) + \epsilon_A \\
&= \Phi(t) + \mathbb{P}\left\{\frac{t}{1-\delta_C} < N(0,1) < \frac{t}{1-\delta_C} + \delta_B\right\} + \mathbb{P}\left\{t < N(0,1) < \frac{t}{1-\delta_C}\right\} + \epsilon_A \ .
\end{aligned}$$

Since the density of the normal distribution is bounded by $\frac{1}{\sqrt{2\pi}} \leqslant 1$,

$$\leqslant \Phi(t) + \delta_B + \mathbb{P}\left\{t < N(0,1) < \frac{t}{1-\delta_C}\right\} + \epsilon_A \ .$$

Applying Lemma D.1 (stated below),

$$\leqslant \Phi(t) + \delta_B + \left(\frac{1}{1-\delta_C} - 1\right) + \epsilon_A \ .$$

Therefore, for all $t \geqslant 0$,

$$\left|\mathbb{P}\left\{(A+\widetilde{B})\cdot(1+\widetilde{C}) \leqslant t\right\} - \Phi(t)\right| \leqslant \delta_B + \frac{\delta_C}{1-\delta_C} + \epsilon_A \ .$$

By identical arguments, we can prove the same for $t \leqslant 0$. □

Lemma D.1. *For any* $0 \leqslant a \leqslant b$,

$$\mathbb{P}\left\{a < N(0,1) < b\right\} \leqslant \left(\frac{b}{a} - 1\right)\cdot\frac{1}{\sqrt{2\pi e}} \leqslant \left(\frac{b}{a}-1\right) \ .$$

Proof.

$$\begin{aligned}\mathbb{P}\left\{a < N(0,1) < b\right\} &= \int_{t=a}^{b} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathsf{d}t\\ &\leqslant (b-a)\cdot\frac{1}{\sqrt{2\pi}}e^{-a^2/2}\\ &= \left(\frac{b}{a}-1\right)\cdot\frac{1}{\sqrt{2\pi}}\cdot a\cdot e^{-a^2/2}\\ &\leqslant \left(\frac{b}{a}-1\right)\cdot\frac{1}{\sqrt{2\pi e}}\ ,\end{aligned}$$

where the last step holds because $\sup_{t>0}\{t\cdot e^{-t^2/2}\} = \frac{1}{\sqrt{e}}$. □

D.2 Sign vector of a transelliptical distribution

Lemma 4.4. *Let*

$$X, X' \overset{iid}{\sim} \mathsf{TE}(\Sigma, \xi; f_1, \dots, f_p) \ .$$

Suppose that Σ *is positive definite, and that* $\xi > 0$ *with probability* 1. *Then* $\mathrm{sign}(X - X')$ *is equal in distribution to* $\mathrm{sign}(Z)$, *where* $Z \sim N(0, \Sigma)$.

Proof of Lemma 4.4. First, since the f_j's are strictly monotone, we see that $\mathrm{sign}(X - X')$ has the same distribution regardless of the choice of the f_j's (assuming without loss of generality that the f_j's are increasing). Therefore it suffices to consider the case that the f_j's are each the identity function, and so $X, X' \overset{iid}{\sim} \mathsf{E}(\mathbf{0}, \Sigma, \xi)$, that is, a zero-mean elliptical distribution. In this case, by Lindskog et al. (2003, Lemma 1), $X - X' \sim \mathsf{E}(\mathbf{0}, \Sigma, \zeta)$ where the distribution of the random variable $\zeta \geqslant 0$ obeys $\varphi_\zeta(t) = \varphi_\xi(t)^2$, where φ_ζ and φ_ξ are the characteristic functions of ζ and ξ, respectively. Note that for two independent copies $\xi_1, \xi_2 \overset{iid}{\sim} \xi$, we have $\varphi_{\xi_1+\xi_2} = \varphi_{\xi_1}\cdot\varphi_{\xi_2} = \varphi_\xi^2 = \varphi_\zeta$, and therefore, $\zeta \overset{\mathcal{D}}{=} \xi_1 + \xi_2$. Since $\xi > 0$ with probability 1, this proves that $\zeta > 0$ with probability 1.

Next take $Z \sim N(0, \Sigma)$. Then $\frac{\Sigma^{-1/2}Z}{||\Sigma^{-1/2}Z||_2}$ is uniformly distributed on the unit sphere, and so

$$\zeta \cdot \Sigma^{1/2} \cdot \frac{\Sigma^{-1/2}Z}{||\Sigma^{-1/2}Z||_2} \sim \mathsf{E}(\mathbf{0}, \Sigma, \zeta) ,$$

which is the distribution of $X - X'$. Using the fact that $\zeta > 0$ with probability 1, we see that $\text{sign}(X - X')$ is equal in distribution to

$$\text{sign}\left(\zeta \cdot \Sigma^{1/2} \cdot \frac{\Sigma^{-1/2}Z}{||\Sigma^{-1/2}Z||_2}\right) = \text{sign}(\Sigma^{1/2} \cdot \Sigma^{-1/2}Z) = \text{sign}(Z) ,$$

as desired. □

D.3 Proof of Lemma 4.7

Lemma 4.7. *The following bound holds deterministically: for any $k \geqslant 1$,*

$$\sup_{u,v \in \mathcal{B}_k} \left| u^\top(\widehat{\Sigma} - \Sigma)v \right| \leqslant \frac{\pi^2}{8} \cdot k||\widehat{T} - T||_\infty^2 + 2\pi \sup_{u \in \mathcal{S}_{k+1}} \left| u^\top(\widehat{T} - T)u \right| .$$

Proof of Lemma 4.7. To prove this theorem, we first state the Transfer Principle of Oliveira (2013):

Lemma D.2 (Lemma 5.1 of Oliveira (2013))**.** *Suppose that $B, C \in \mathbb{R}^{p \times p}$ are matrices with non-negative diagonals, satisfying*

$$v^\top Bv \geqslant v^\top Cv \cdot (1 - \eta) \textit{ for all } (k+1)\textit{-sparse } v \in \mathbb{R}^p.$$

Let $d_i = B_{ii} - (1-\eta)C_{ii} \geqslant 0$. Then

$$v^\top Bv \geqslant v^\top Cv \cdot (1 - \eta) - \frac{||\text{diag}\{\sqrt{d}\} \cdot v||_1^2}{k} \textit{ for all } v \in \mathbb{R}^p .$$

Now we turn to proving the theorem. By Taylor's theorem,

$$\widehat{\Sigma} = \Sigma + \frac{\pi}{2}\cos\left(\frac{\pi}{2}T\right) \circ \left(\widehat{T} - T\right) - \frac{\pi^2}{8}\sin\left(\frac{\pi}{2}\bar{T}\right) \circ \left(\widehat{T} - T\right) \circ \left(\widehat{T} - T\right)$$

where $\bar{T}$ has entries $\bar{\tau}_{ab} = (1 - t_{ab})\tau_{ab} + t_{ab}\widehat{\tau}_{ab}$, with $t_{ab} \in [0, 1]$ for each a, b. Taking any $u, v \in \mathcal{B}_k$, then,

$$\left|u^\top(\widehat{\Sigma} - \Sigma)v\right| \leqslant \frac{\pi}{2}\left|u^\top\left[\cos\left(\frac{\pi}{2}T\right) \circ \left(\widehat{T} - T\right)\right]v\right| + \frac{\pi^2}{8}\left|u^\top\left[\sin\left(\frac{\pi}{2}\bar{T}\right) \circ \left(\widehat{T} - T\right) \circ \left(\widehat{T} - T\right)\right]v\right| .$$

First, to bound the $\sin(\cdot)$ matrix term, note that

$$\begin{aligned}&\left|u^\top\left[\sin\left(\frac{\pi}{2}\bar{T}\right) \circ \left(\widehat{T} - T\right) \circ \left(\widehat{T} - T\right)\right]v\right| \\ &\qquad \leqslant ||u||_1||v||_1\left\|\sin\left(\frac{\pi}{2}\bar{T}\right) \circ \left(\widehat{T} - T\right) \circ \left(\widehat{T} - T\right)\right\|_\infty \leqslant ||u||_1||v||_1||\widehat{T} - T||_\infty^2 ,\end{aligned}$$

where the last step holds since the $\sin(\cdot)$ function lies in $[-1,1]$. Furthermore, $||u||_1, ||v||_1 \leqslant \sqrt{k}$ for all $u, v \in \mathcal{B}_k$ by definition.

Next, we bound the $\cos(\cdot)$ matrix term. By Lemma C.1, we can express $\cos\left(\frac{\pi}{2}T\right)$ as a convex combination,

$$\cos\left(\frac{\pi}{2}T\right) = \sum_r t_r \cdot a_r b_r^\top ,$$

where $a_r, b_r \in \mathbb{R}^p$ satisfy $||a_r||_\infty, ||b_r||_\infty \leqslant 1$ for all r, and $t_r \geqslant 0$ satisfy $\sum_r t_r = 4$. Furthermore, for $u, v \in \mathcal{B}_k$ and for each r, note that $u \circ a_r, v \circ b_r \in \mathcal{B}_k$ due to the bound on $||a_r||_\infty, ||b_r||_\infty$. Then

$$\begin{aligned}\left|u^\top \left[\cos\left(\frac{\pi}{2}T\right) \circ \left(\widehat{T} - T\right)\right] v\right| &\leqslant \sum_r t_r \left|u^\top \left[a_r b_r^\top \circ \left(\widehat{T} - T\right)\right] v\right| \\ &= \sum_r t_r \left|(u \circ a_r)^\top \left(\widehat{T} - T\right)(v \circ b_r)\right| \leqslant 4 \sup_{u',v' \in \mathcal{B}_k} \left|u'^\top (\widehat{T} - T) v'\right| .\end{aligned}$$

Finally, to reduce to the sparse set $\mathcal{S}_{k+1}$, we use Oliviera's Transfer Principle (Lemma D.2). Define $B = \mathbf{I} - (\widehat{T} - T)$ and $C = \mathbf{I}$, and let

$$\eta = \sup_{u \in \mathcal{S}_{k+1}} \left|u^\top (\widehat{T} - T) u\right| .$$

Then, for all $(k+1)$-sparse vectors x, by considering the rescaled vector $u = x/||x||_2$, we see that

$$x^\top B x = ||x||_2^2 - x^\top (\widehat{T} - T) x \geqslant (1 - \eta)||x||_2^2 = (1 - \eta) \cdot x^\top C x .$$

Furthermore, for each $i = 1, \ldots, p$, we have $C_{ii} = 1$ trivially and $B_{ii} = 1 - \widehat{T}_{ii} + T_{ii} = 1$, which is true because $\widehat{T}_{ii} = T_{ii} = 1$ by definition. Then in the notation of Lemma D.2, for each $i = 1, \ldots, p$ we set $d_i = B_{ii} - (1 - \eta)C_{ii} = \eta$. Applying Lemma D.2, then, for all $x \in \mathbb{R}^p$,

$$x^\top B x \geqslant (1 - \eta) x^\top C x - \eta \frac{||x||_1^2}{k} ,$$

and plugging in our definitions of B and C, we get

$$x^\top (\widehat{T} - T) x \leqslant \eta \left(||x||_2^2 + \frac{||x||_1^2}{k}\right) .$$

By symmetry, we can instead set $B = \mathbf{I} + (\widehat{T} - T)$ to obtain the same upper bound on $-x^\top (\widehat{T} - T) x$.

To conclude, take any $u, v \in \mathcal{B}_k$. Then, setting $x = \frac{u+v}{2}$ and $y = \frac{u-v}{2}$, observe that $x, y \in \mathcal{B}_k$ also, and that

$$\left|u^\top (\widehat{T} - T) v\right| = \frac{1}{2}\left|x^\top (\widehat{T} - T) x - y^\top (\widehat{T} - T) y\right| \leqslant \eta ,$$

which proves that

$$\sup_{u,v \in \mathcal{B}_k} \left|u^\top (\widehat{T} - T) v\right| \leqslant \eta = \sup_{u \in \mathcal{S}_{k+1}} \left|u^\top (\widehat{T} - T) u\right| ,$$

and thus we obtain the desired result. □

D.4 Proof of Lemma 4.6

Lemma 4.6. *Suppose that $k \geqslant 1$ and $\delta \in (0,1)$ satisfy $\log(2/\delta) + k\log(12p) \leqslant n$. Then with probability at least $1-\delta$ it holds that*

$$\sup_{u\in\mathcal{S}_k} \left| u^\top (\widehat{T} - T) u \right| \leqslant 16(1+\sqrt{5})\mathsf{C}(\Sigma)\cdot\sqrt{\frac{\log(2/\delta)+k\log(12p)}{n}} .$$

Proof of Lemma 4.6. This lemma is a straightforward combination of Lemma C.2 (stated in Appendix C) together with the following result:

Lemma D.3 (Adapted from Lemma 5.1 and Theorem 5.2 of Baraniuk et al. (2008))**.** *Let A be a random matrix satisfying*

$$\exp\left\{t\cdot u^\top A u\right\} \leqslant \exp\left\{\frac{c_1 t^2}{n}\right\} \text{ for all } |t| \leqslant c_0 n \text{ and all unit vectors } u\in\mathbb{R}^p \tag{D.1}$$

for some constants c_0, c_1. Then for any $k \geqslant 1$ and any $\delta\in(0,1)$ satisfying

$$\log(2/\delta) + k\log(12p) \leqslant nc_0^2 c_1 ,$$

with probability at least $1-\delta$ it holds that

$$|u^\top A u| \leqslant \sqrt{\frac{16c_1}{n}\left(\log(2/\delta)+k\log(12p)\right)} \text{ for all } k\text{-sparse unit vectors } u\in\mathbb{R}^p. \tag{D.2}$$

Combined, Lemmas D.3 and C.2 immediately yield Lemma 4.6, as desired. □

We next turn to the proof of Lemma D.3.

Proof of Lemma D.3. (Adapted from Lemma 5.1 and Theorem 5.2 of Baraniuk et al. (2008).) First fix any $S\subset[p]$ with $|S| = k$. Let $\epsilon = \sqrt{\frac{16c_1}{n}\left(\log(2/\delta)+k\log(12p)\right)}$. Following the same arguments as in Baraniuk et al. (2008, Lemma 5.1), we can take a set $\mathcal{U}\subset\mathbb{R}^S$ of unit vectors, with $|\mathcal{U}| \leqslant 12^k$, such that

$$\sup_{\text{unit } u\in\mathbb{R}^S} \left|u^\top A u\right| \leqslant 2\sup_{\widetilde{u}\in\mathcal{U}} \left|\widetilde{u}^\top A\widetilde{u}\right| .$$

Furthermore, for any fixed $\widetilde{u}\in\mathcal{U}$, for any $0 < t \leqslant c_0 n$,

$$\begin{aligned}\mathbb{P}\left\{\widetilde{u}^\top A\widetilde{u} > \epsilon/2\right\} &\leqslant \mathbb{E}\left[t\cdot\widetilde{u}^\top A\widetilde{u} - t\cdot\epsilon/2\right]\\ &\leqslant \exp\left(\frac{c_1 t^2}{n} - t\cdot\epsilon/2\right) .\end{aligned}$$

Setting $t = \frac{n\epsilon}{4c_1} \leqslant c_0 n$,

$$= \exp\left(-\frac{n\epsilon^2}{16c_1}\right) ,$$

and similarly,

$$\mathbb{P}\left\{\widetilde{u}^\top A\widetilde{u} < -\epsilon/2\right\} \leqslant \exp\left(-\frac{n\epsilon^2}{16c_1}\right) .$$

Therefore,

$$\mathbb{P}\left\{\sup_{\widetilde{u}\in\mathcal{U}}\left|\widetilde{u}^\top A\widetilde{u}\right| > \epsilon/2\right\} \leqslant 2\cdot 12^k\cdot\exp\left(-\frac{n\epsilon^2}{16c_1}\right) \ ,$$

and so

$$\mathbb{P}\left\{\sup_{\text{unit } u\in\mathbb{R}^S}\left|u^\top Au\right| > \epsilon\right\} \leqslant 2\cdot 12^k\cdot\exp\left(-\frac{n\epsilon^2}{16c_1}\right) \ .$$

Finally, taking all $\binom{p}{k} \leqslant p^k$ choices for S, we see that

$$\mathbb{P}\left\{\sup_{k\text{-sparse unit } u,v\,\in\,\mathbb{R}^p}\left\{u^\top Au\right\} \leqslant \epsilon\right\} \geqslant 1 - 2(12p)^k\cdot\exp\left(-\frac{n\epsilon^2}{16c_1}\right) \ .$$

□

D.5 Proof of Lemma B.4

Lemma B.4. *The following bound holds deterministically:*

$$||\breve{\Theta} - \widetilde{\Theta}||_\infty \leqslant \mathsf{C}(\Sigma)\cdot\max_{c\in\{a,b\}}||\Delta_c||_2^2 +$$
$$\sup_{u,v\in\mathcal{B}_k}\left|u^\top(\widehat{\Sigma}-\Sigma)v\right|\cdot\left(2\max_{c\in\{a,b\}}||\Delta_c||_{(k)}\cdot\left(2+\max_{c\in\{a,b\}}||\gamma_c||_{(k)}\right) + \max_{c\in\{a,b\}}||\Delta_c||^2_{(k)}\right) \ .$$

(The definition of the norm $||\cdot||_{(k)}$ is given in (B.6).)

Proof of Lemma B.4. Choose any $c, d \in \{a, b\}$; we will bound the (c, d)th entry of the error, that is, $\left|\breve{\Theta}_{cd} - \widetilde{\Theta}_{cd}\right|$. Write $\Delta_c = \breve{\gamma}_c - \gamma_c$ for each $c = a, b$. We have

$$\begin{aligned}
&\left|\breve{\Theta}_{cd} - \widetilde{\Theta}_{cd}\right| \\
&= \left|\left(\widehat{\Sigma}_{cd} - \breve{\gamma}_c^\top\widehat{\Sigma}_{Id} - \widehat{\Sigma}_{Ic}^\top\breve{\gamma}_d + \breve{\gamma}_c^\top\widehat{\Sigma}_I\breve{\gamma}_d\right) - \left(\widehat{\Sigma}_{cd} - \gamma_c^\top\widehat{\Sigma}_{Id} - \widehat{\Sigma}_{Ic}^\top\gamma_d + \gamma_c^\top\widehat{\Sigma}_I\gamma_d\right)\right| \\
&= \left|\breve{\gamma}_c^\top\widehat{\Sigma}_I\breve{\gamma}_d - \gamma_c^\top\widehat{\Sigma}_I\gamma_d - \Delta_c^\top\widehat{\Sigma}_{Id} - \widehat{\Sigma}_{Ic}^\top\Delta_d\right| \\
&= \left|\Delta_c^\top\widehat{\Sigma}_I\gamma_d + \gamma_c^\top\widehat{\Sigma}_I\Delta_d + \Delta_c^\top\widehat{\Sigma}_I\Delta_d - \Delta_c^\top\widehat{\Sigma}_{Id} - \widehat{\Sigma}_{Ic}^\top\Delta_d\right| \\
&\leqslant \left|\Delta_c^\top\Sigma_I\gamma_d + \gamma_c^\top\Sigma_I\Delta_d + \Delta_c^\top\Sigma_I\Delta_d - \Delta_c^\top\Sigma_{Id} - \Sigma_{Ic}^\top\Delta_d\right| \qquad (D.3) \\
&\qquad + \left|\Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\gamma_d + \gamma_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d + \Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d - \Delta_c^\top(\widehat{\Sigma}_{Id} - \Sigma_{Id}) - (\widehat{\Sigma}_{Ic} - \Sigma_{Ic})^\top\Delta_d\right| \ .
\end{aligned}$$

Now we bound each of these terms. To bound the first term on the right-hand side of (D.3), we have

$$\begin{aligned}
&\left|\Delta_c^\top\Sigma_I\gamma_d + \gamma_c^\top\Sigma_I\Delta_d + \Delta_c^\top\Sigma_I\Delta_d - \Delta_c^\top\Sigma_{Id} - \Sigma_{Ic}^\top\Delta_d\right| \\
&= \left|\Delta_c^\top\Sigma_I\Sigma_I^{-1}\Sigma_{Id} + \Sigma_{Ic}^\top\Sigma_I^{-1}\Sigma_I\Delta_d + \Delta_c^\top\Sigma_I\Delta_d - \Delta_c^\top\Sigma_{Id} - \Sigma_{Ic}^\top\Delta_d\right| \\
&= \left|\Delta_c^\top\Sigma_I\Delta_d\right| \\
&\leqslant ||\Delta_c||_2\cdot||\Delta_d||_2\cdot||\Sigma||_{\mathsf{op}} \\
&\leqslant ||\Delta_c||_2\cdot||\Delta_d||_2\cdot\mathsf{C}(\Sigma) \ ,
\end{aligned}$$

where the last step holds because

$$||\Sigma_I|| \leqslant ||\Sigma|| = \lambda_{\min}(\Sigma) \cdot \mathsf{C}(\Sigma) \ ,$$

and we must have $\lambda_{\min}(\Sigma) \leqslant 1$ because $\mathrm{diag}(\Sigma) = \mathbf{1}$.

Finally, to bound the second term on the right-hand side of (D.3), we have

$$\begin{aligned}
&\left|\Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\gamma_d + \gamma_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d + \Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d - \Delta_c^\top(\widehat{\Sigma}_{Id} - \Sigma_{Id}) - (\widehat{\Sigma}_{Ic} - \Sigma_{Ic})^\top\Delta_d\right| \\
&\leqslant \left|\Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\gamma_d\right| + \left|\gamma_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d\right| + \left|\Delta_c^\top(\widehat{\Sigma}_I - \Sigma_I)\Delta_d\right| + \left|\Delta_c^\top(\widehat{\Sigma}_{Id} - \Sigma_{Id})\right| + \left|(\widehat{\Sigma}_{Ic} - \Sigma_{Ic})^\top\Delta_d\right| \\
&\leqslant \sup_{u,v\in\mathcal{B}_k} \left|u^\top(\widehat{\Sigma} - \Sigma)v\right| \cdot \\
&\qquad \left(||\Delta_c||_{(k)}||\gamma_d||_{(k)} + ||\gamma_c||_{(k)}||\Delta_d||_{(k)} + ||\Delta_c||_{(k)}||\Delta_d||_{(k)} + ||\Delta_c||_{(k)}||\mathbf{e}_d||_{(k)} + ||\mathbf{e}_c||_{(k)}||\Delta_d||_{(k)}\right) \ ,
\end{aligned}$$

where $\mathbf{e}_c$ and $\mathbf{e}_d$ are the cth and dth basis vectors in $\mathbb{R}^p$. Since $||\mathbf{e}_c||_{(k)} = ||\mathbf{e}_d||_{(k)} \leqslant 2$, the desired result of the lemma follows trivially from these bounds. □

D.6 Proof of Lemma B.1

Lemma B.1. *Suppose that Assumptions 3.1, 3.2 and 3.4 hold. Let $g(X, X')$ and $g_1(X)$ be defined as in the proof of Theorem 4.1. Then*

$$\nu_{g_1}^2 := \mathsf{Var}(g_1(X)) \geqslant \frac{1}{\pi^2} C_{\mathsf{variance}}^2$$

and

$$\nu_{g_1}^3 \leqslant \eta_g^3 := \mathbb{E}\left[|g(X, X')|^3\right] \leqslant C_{\mathsf{moment}}$$

where $C_{\mathsf{variance}}, C_{\mathsf{moment}}$ are constants depending only on $C_{\mathsf{cov}}, C_{\mathsf{kernel}}$ and not on (n, p_n, k_n).

Proof of Lemma B.1. First, we have

$$\begin{aligned}
g_1(X) =& \mathbb{E}\left[\mathrm{sign}(X - X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mathrm{sign}(X - X') \mid X\right] \\
&= \mathbb{E}\left[\left(\mathrm{sign}(X - X') \otimes \mathrm{sign}(X - X')\right)^\top \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mid X\right] \\
&= \mathbb{E}\left[\left(\mathrm{sign}(X - X') \otimes \mathrm{sign}(X - X')\right) \mid X\right]^\top \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \\
&= h_1(X)^\top \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \ ,
\end{aligned}$$

where $h_1(X)$ is defined in Assumption 3.4, and has variance Σ_H. Therefore,

$$\begin{aligned}
\nu_{g_1}^2 &= \mathsf{Var}(g_1(X)) \\
&= \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top \cdot \Sigma_{h_1} \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \ .
\end{aligned}$$

By Assumption 3.4,

$$\begin{aligned}
&\geqslant C_{\mathsf{kernel}} \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top \cdot \Sigma_h \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \\
&= C_{\mathsf{kernel}} \cdot \mathsf{Var}\left(\mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top h(X,X')\right) \\
&= C_{\mathsf{kernel}} \cdot \mathsf{Var}\left(\mathrm{sign}(X-X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mathrm{sign}(X-X')\right) \ .
\end{aligned}$$

For $Z \sim N(0,\Sigma)$, applying Lemma 4.4, $\mathrm{sign}(X-X')$ has the same distribution as $\mathrm{sign}(Z)$,

$$\begin{aligned}
&= C_{\mathsf{kernel}} \cdot \mathsf{Var}\left(\mathrm{sign}(Z)^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mathrm{sign}(Z)\right) \\
&\geqslant C_{\mathsf{kernel}} \cdot C_{\mathsf{signs}} \cdot \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)_{ab}^2 \ ,
\end{aligned}$$

where the last step applies the following lemma (proved in Appendix D.8).

Lemma D.4. *Take any positive definite $\Sigma \in \mathbb{R}^{p\times p}$, any distinct $a, b \in \{1,\dots,p\}$, and any matrix $M \in \mathbb{R}^{p\times p}$ with $M_{ja} = 0$ for all j. Let $Z \sim N(0,\Sigma)$. Then there exists a constant C_{signs} depending on $\mathsf{C}(\Sigma)$ only, such that*

$$\mathsf{Var}\left(\mathrm{sign}(Z)^\top M \,\mathrm{sign}(Z)\right) \geqslant C_{\mathsf{signs}} \cdot M_{ab}^2 \ .$$

Finally, we have

$$\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)_{ab}^2 = u_a^2 v_b^2 \cos\left(\frac{\pi}{2}T_{ab}\right)^2 \geqslant (C_{\mathsf{cov}})^{-2} \ ,$$

where the last step holds because $u_a = v_b = 1$ and

$$\cos\left(\frac{\pi}{2}T_{ab}\right) = \sqrt{1 - \sin\left(\frac{\pi}{2}T_{ab}\right)^2} = \sqrt{1-\Sigma_{ab}^2} \geqslant 1 - \Sigma_{ab} = \lambda_{\min}\left(\Sigma_{ab,ab}\right) \geqslant (C_{\mathsf{cov}})^{-1} \ .$$

To summarize, we have

$$\nu_{g_1}^2 \geqslant \frac{C_{\mathsf{kernel}} C_{\mathsf{signs}}}{C_{\mathsf{cov}}^2} =: \frac{1}{\pi^2} C_{\mathsf{variance}}^2 \ .$$

Next, we give an upper bound on $\nu_{g_1}^2$:

$$\begin{aligned}
\nu_{g_1}^2 &= \mathsf{Var}(g_1(X)) \\
&= \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top \cdot \Sigma_{h_1} \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \\
&\leqslant \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top \cdot \Sigma_{h} \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \ .
\end{aligned}$$

As for the lower bound,

$$\begin{aligned}
&= \mathsf{Var}\left(\mathrm{sign}(X-X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mathrm{sign}(X-X')\right) \\
&= \mathsf{Var}\left(g(X,X')\right) \leqslant \mathbb{E}\left[|g(X,X')|^2\right] \leqslant \mathbb{E}\left[|g(X,X')|^3\right]^{2/3} = \eta_g^2 \ .
\end{aligned}$$

Finally, we compute an upper bound on $\eta_g^3 = \mathbb{E}\left[|g(X,X')|^3\right]$. By Lemma C.1, there exists a decomposition

$$\cos\left(\frac{\pi}{2}T\right) = \sum_r t_r a_r b_r^\top$$

where $t_r \geqslant 0$, $\sum_r t_r \leqslant 4$, and $||a_r||_\infty, ||b_r||_\infty \leqslant 1$. Note that, by (B.1),

$$||u||_2 = \sqrt{1 + ||\gamma_a||_2^2} \leqslant \sqrt{1 + C_{\textsf{cov}}^2}$$

and similarly $||v||_2 \leqslant \sqrt{1 + C_{\textsf{cov}}^2}$. Then for each r,

$$||u \circ a_r||_2 \vee ||v \circ b_r||_2 \leqslant \sqrt{1 + C_{\textsf{cov}}^2}\ .$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[|g(X,X')|^3\right] &= \mathbb{E}\left[\left|\text{sign}(X-X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)\text{sign}(X-X')\right|^3\right] \\
&= \mathbb{E}\left[\left|\sum_r t_r \cdot \text{sign}(X-X')^\top \left(uv^\top \circ a_r b_r^\top\right)\text{sign}(X-X')\right|^3\right] \\
&\leqslant \sum_r t_r \cdot \mathbb{E}\left[\left|\text{sign}(X-X')^\top \left(uv^\top \circ a_r b_r^\top\right)\text{sign}(X-X')\right|^3\right] \\
&\leqslant 4 \cdot \max_r \mathbb{E}\left[\left|\text{sign}(X-X')^\top \left(uv^\top \circ a_r b_r^\top\right)\text{sign}(X-X')\right|^3\right] \\
&= 4 \cdot \max_r \mathbb{E}\left[\left|\text{sign}(X-X')^\top(u \circ a_r)\right|^3 \cdot \left|\text{sign}(X-X')^\top(v \circ b_r)\right|^3\right] \\
&= 4 \cdot \max_r \sqrt{\mathbb{E}\left[\left|\text{sign}(X-X')^\top(u \circ a_r)\right|^6\right]} \cdot \sqrt{\mathbb{E}\left[\left|\text{sign}(X-X')^\top(v \circ b_r)\right|^6\right]} \\
&\leqslant 4||u \circ a_r||_2^3 \cdot ||v \circ b_r||_2^3 \cdot \max_{||w||_2=1} \mathbb{E}\left[\left|\text{sign}(X-X')^\top w\right|^6\right] \\
&\leqslant 4(1 + C_{\textsf{cov}}^2)^3 \cdot \max_{||w||_2=1} \mathbb{E}\left[\left|\text{sign}(X-X')^\top w\right|^6\right] \\
&\leqslant 4(1 + C_{\textsf{cov}}^2)^3 \cdot C_{\textsf{cov}}^3 \cdot 6! \cdot 2\sqrt{e} =: C_{\textsf{moment}}\ ,
\end{aligned}$$

where the last step holds because $\text{sign}(X-X')$ is $C_{\textsf{cov}}$-subgaussian by Lemmas 4.4 and 4.5. □

D.7 Proofs of lemmas for the initial estimators

Lemma C.1. *There exist vectors $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ with $||a_r||_\infty, ||b_r||_\infty \leqslant 1$ for all $r \geqslant 1$, and a sequence $t_1, t_2, \cdots \geqslant 0$ with $\sum_r t_r = 4$, such that $\cos\left(\frac{\pi}{2}T\right) = \sum_{r \geqslant 1} t_r a_r b_r^\top$.*

Proof of Lemma C.1. We will use the matrix max norm, defined for a matrix $M \in \mathbb{R}^{d_1 \times d_2}$ as

$$||M||_{\max} = \min\left\{\max_{1 \leqslant i \leqslant d_1} ||A_{(i)}||_2 \cdot \max_{1 \leqslant j \leqslant d_2} ||B_{(j)}||_2 \ :\ r \geqslant 1, A \in \mathbb{R}^{d_1 \times r}, B \in \mathbb{R}^{d_2 \times r} \text{ s.t. } M = A \cdot B^\top\right\}\ ,$$

where $A_{(i)}$ and $B_{(j)}$ denote the ith row of A and the jth row of B, respectively. The matrix max norm satisfies several key properties that we will use here (Srebro and Shraibman, 2005): first,

$$W \succeq 0 \ \Rightarrow ||W||_{\max} \leqslant \max_i W_{ii}\ ; \tag{D.4}$$

second,

$$||W||_{\max} \leqslant 1 \;\Rightarrow\; \frac{W}{2} \in \mathsf{ConvexHull}\left\{ab^\top : ||a||_\infty, ||b||_\infty \leqslant 1\right\} \;; \tag{D.5}$$

and finally,

$$||W \circ (uv^\top)||_* \leqslant ||W||_{\max} \text{ for all unit vectors } u, v \text{ and all matrices } W \;, \tag{D.6}$$

where recall that $||\cdot||_*$ is the matrix nuclear norm (the sum of the singular values).

For our matrix $\cos\left(\frac{\pi}{2}T\right)$, Wegkamp and Zhao (2013) show that

$$\cos\left(\frac{\pi}{2}T\right) = \sum_{r\geqslant 0} \binom{1/2}{r} (-1)^r \Sigma \circ_{2r} \Sigma, \text{ and } \Sigma \circ_{2r} \Sigma \succeq \mathbf{0} \text{ for all } r \;,$$

where $\Sigma \circ_{2r} \Sigma$ is the matrix with entries given by elementwise powers of Σ, that is, $(\Sigma \circ_{2r} \Sigma)_{jk} = (\Sigma_{jk})^{2r}$. Then for each $r \geqslant 0$, applying (D.4),

$$||\Sigma \circ_{2r} \Sigma||_{\max} \leqslant \max_i (\Sigma \circ_{2r} \Sigma)_{ii} = \max_i (\Sigma_{ii})^{2r} = 1 \;,$$

since Σ is a correlation matrix. Then

$$\begin{aligned} ||\cos\left(\frac{\pi}{2}T\right)||_{\max} &= ||\sum_{r\geqslant 0} \binom{1/2}{r} (-1)^r \Sigma \circ_{2r} \Sigma||_{\max} \\ &\leqslant \sum_{r\geqslant 0} \left|\binom{1/2}{r}\right| \cdot ||\Sigma \circ_{2r} \Sigma||_{\max} \leqslant \sum_{r\geqslant 0} \left|\binom{1/2}{r}\right| = 2 \;, \end{aligned}$$

where the last identity comes from Wegkamp and Zhao (2013). Finally, by (D.5), we have

$$\frac{\cos\left(\frac{\pi}{2}T\right)}{4} \in \mathsf{ConvexHull}\left\{ab^\top : ||a||_\infty, ||b||_\infty \leqslant 1\right\}$$

and so $\cos\left(\frac{\pi}{2}T\right)$ can be expressed as a convex combination as stated in the lemma. □

Lemma C.2. *For fixed u, v with $||u||_2, ||v||_2 \leqslant 1$, for any $|t| \leqslant \frac{n}{4(1+\sqrt{5})C_{\mathsf{cov}}}$,*

$$\mathbb{E}\left[\exp\left(t \cdot u^\top (\widehat{T} - T) v\right)\right] \leqslant \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 t^2 \cdot C_{\mathsf{cov}}^2}{n}\right) \;.$$

Proof of Lemma C.2. We start with a simple observation that

$$u^T\left(\widehat{T} - T\right)v = \frac{1}{4}(u+v)^\top(\widehat{T} - T)(u+v) - \frac{1}{4}(u-v)^\top(\widehat{T} - T)(u-v),$$

which gives us (via Cauchy-Schwartz)

$$\begin{aligned} &\mathbb{E}\left[\exp\left(t \cdot u^\top (\widehat{T} - T) v\right)\right] \\ &= \mathbb{E}\left[\exp\left(t \cdot \frac{1}{4}(u+v)^\top(\widehat{T} - T)(u+v) - t \cdot \frac{1}{4}(u-v)^\top(\widehat{T} - T)(u-v)\right)\right] \\ &\leqslant \sqrt{\mathbb{E}\left[\exp\left(t \cdot \frac{1}{2}(u+v)^\top(\widehat{T} - T)(u+v)\right)\right]} \cdot \sqrt{\mathbb{E}\left[\exp\left(-t \cdot \frac{1}{2}(u-v)^\top(\widehat{T} - T)(u-v)\right)\right]} \;. \end{aligned}$$

Note that $||\frac{1}{2}(u+v)||_2 \vee ||\frac{1}{2}(u-v)||_2 \leqslant 1$. Therefore, it will be sufficient to show that for any $|t| \leqslant \frac{n}{8C_{\mathsf{cov}}}$ and any unit vector w,

$$\mathbb{E}\left[\exp\left(2t \cdot w^\top(\widehat{T} - T)w\right)\right] \leqslant \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 t^2 \cdot C_{\mathsf{cov}}^2}{n}\right) . \tag{D.7}$$

We will prove (D.7) using the Chernoff bounding technique. To that end, denote S_n the group of permutations of $[n]$, and for any i, let $X_{(i)}$ denote the i-th row of X. For a fixed $w \in \mathbb{R}^p$ and $\sigma \in S_n$, define

$$Z_{\sigma,i} = w^\top \left(\operatorname{sign}\left((X_{(\sigma(i))} - X_{(\sigma(i+n/2))})(X_{(\sigma(i))} - X_{(\sigma(i+n/2))})^\top\right) - T\right) w \ .$$

Observe that

$$w^T \left(\widehat{T} - T\right) w = \frac{1}{n!} \sum_{\sigma \in S_n} \frac{2}{n} \sum_{i \in [n/2]} Z_{\sigma,i}, \tag{D.8}$$

and that for any fixed $\sigma \in S_n$, the $Z_{\sigma,i}$'s are i.i.d. for $i = 1, \ldots, n/2$, and are identically distributed as

$$\widetilde{Z} = w^\top \left(\operatorname{sign}\left((X_{(i)} - X_{(i+n/2)})(X_{(i)} - X_{(i+n/2)})^\top\right) - T\right) w \ .$$

Using Lemma 4.4 and Lemma 4.5, for any fixed unit vector $w \in \mathbb{R}^p$, $w^T \operatorname{sign}\left(X_{(i)} - X_{(i+n/2)}\right)$ is a C_{cov}-subgaussian random variable, and

$$\widetilde{Z} = \left(w^T \operatorname{sign}\left(X_{(i)} - X_{(i+n/2)}\right)\right)^2 - \mathbb{E}\left[\left(w^T \operatorname{sign}\left(X_{(i)} - X_{(i+n/2)}\right)\right)^2\right] .$$

Applying Lemma D.5 (stated below), for any $|t| \leqslant \frac{1}{2(1+\sqrt{5})C_{\mathsf{cov}}}$,

$$\mathbb{E}\left[\exp\left(t\widetilde{Z}\right)\right] \leqslant \exp\left(\frac{32t^2 C_{\mathsf{cov}}^2}{1 - 4tC_{\mathsf{cov}}}\right) \leqslant \exp\left(\frac{32t^2 C_{\mathsf{cov}}^2}{1 - \frac{2}{1+\sqrt{5}}}\right) \leqslant \exp\left(8(1+\sqrt{5})^2 t^2 C_{\mathsf{cov}}^2\right) .$$

Then, referring back to (D.8), for $0 < t \leqslant \frac{n}{4(1+\sqrt{5})C_{\mathsf{cov}}}$,

$$\mathbb{E}\left[\exp\left(tw^T\left(\widehat{T} - T\right)w\right)\right] = \mathbb{E}\left[\exp\left(\frac{t}{n!}\sum_{\sigma \in S_n} \frac{2}{n} \sum_{i \in [n/2]} Z_{\sigma,i}\right)\right] .$$

By Jensen's inequality,

$$\leqslant \frac{1}{n!} \sum_{\sigma \in S_n} \mathbb{E}\left[\exp\left(\frac{2t}{n} \sum_{i \in [n/2]} Z_{\sigma,i}\right)\right] .$$

Since for any fixed σ, the $Z_{\sigma,i}$'s are i.i.d., and are each equal to $\widetilde{Z}$ in distribution,

$$\begin{aligned}
&= \frac{1}{n!}\sum_{\sigma\in S_n}\left(\mathbb{E}\left[\exp\left(\frac{2t}{n}\widetilde{Z}\right)\right]\right)^{n/2} \\
&= \left(\mathbb{E}\left[\exp\left(\frac{2t}{n}\widetilde{Z}\right)\right]\right)^{n/2} \\
&\leqslant \left(\exp\left(8(1+\sqrt{5})^2(2t/n)^2 C_{\mathsf{cov}}^2\right)\right)^{n/2} \\
&= \exp\left(\frac{\left[4(1+\sqrt{5})\right]^2 t^2\cdot C_{\mathsf{cov}}^2}{n}\right) .
\end{aligned}$$

□

Lemma D.5. *Suppose Z is C-subgaussian, that is, $\mathbb{E}\left[\exp\left(tZ\right)\right] \leqslant e^{Ct^2/2}$ for all $t \in \mathbb{R}$. Then*

$$\mathbb{E}\left[\exp\left\{t(Z^2-\mathbb{E}[Z^2])\right\}\right] \leqslant \exp\left(\frac{32t^2C^2}{1-4|t|C}\right)$$

for all $|t| < \frac{1}{4C}$.

We remark that it is well known that the square of a subgaussian random variable satisfies subgaussian tails near to its mean (see, for example, Lemmas 5.5, 5.14, 5.15 in Vershynin, 2012), but here we obtain small explicit constants.

Proof. The first part of this proof follows the arguments in Vershynin (2012, Lemma 5.5). First, we bound $\mathbb{E}\left[Z^{2k}\right]$ for all integers $k \geqslant 1$. We have

$$\mathbb{E}\left[Z^{2k}\right] = \frac{C^k}{(2k)^k}\mathbb{E}\left[\left(\sqrt{\frac{2k}{C}}\cdot Z\right)^{2k}\right] \leqslant \frac{C^k}{(2k)^k}\mathbb{E}\left[(2k)!\cdot\exp\left\{\sqrt{\frac{2k}{C}}\cdot Z\right\}\right]$$
$$\leqslant \frac{C^k}{(2k)^k}(2k)!\cdot\exp\left\{\left(\sqrt{\frac{2k}{C}}\right)^2\cdot C/2\right\} = \frac{C^k(2k)!e^k}{(2k)^k} .$$

Then, for any $t > 0$,

$$\mathbb{E}\left[e^{tZ^2}\right] = 1 + t\mathbb{E}\left[Z^2\right] + \sum_{k\geqslant 2}\mathbb{E}\left[\frac{\left(tZ^2\right)^k}{k!}\right] \leqslant 1 + t\mathbb{E}\left[Z^2\right] + \sum_{k\geqslant 2}\frac{t^kC^ke^k}{(2k)^k}\cdot\frac{(2k)!}{k!} .$$

Using Stirling's approximation to give upper and lower bounds on $(2k)!$ and $k!$, respectively, for each $k \geqslant 2$,

$$\begin{aligned}
&\leqslant 1 + t\mathbb{E}\left[Z^2\right] + \sum_{k\geqslant 2}\frac{t^kC^ke^k}{(2k)^k}\cdot\frac{e\cdot(2k)^{2k+1/2}\cdot e^{-2k}}{\sqrt{2\pi}\cdot k^{k+1/2}\cdot e^{-k}} \\
&= 1 + t\mathbb{E}\left[Z^2\right] + \frac{e}{\sqrt{\pi}}\sum_{k\geqslant 2}(2tC)^k \\
&= 1 + t\mathbb{E}\left[Z^2\right] + \frac{e}{\sqrt{\pi}}\cdot\frac{4t^2C^2}{1-2tC} ,
\end{aligned}$$

as long as $2tC < 1$. Next, trivially, for any $k \geqslant 2$, $\left|\mathbb{E}\left[\left(Z^2 - \mathbb{E}[Z^2]\right)^k\right]\right| \leqslant 2^k \mathbb{E}\left[Z^{2k}\right]$. Then we have, for $|t| < \frac{1}{4C}$,

$$\mathbb{E}\left[\exp\left(t(Z^2 - \mathbb{E}[Z^2])\right)\right] = 1 + \sum_{k \geqslant 2} \frac{t^k}{k!}\mathbb{E}\left[\left(Z^2 - \mathbb{E}[Z^2]\right)^k\right] \leqslant 1 + \sum_{k \geqslant 2} \frac{|t|^k}{k!}\left|\mathbb{E}\left[\left(Z^2 - \mathbb{E}[Z^2]\right)^k\right]\right|$$

$$\leqslant 1 + \sum_{k \geqslant 2} \frac{2^k |t|^k}{k!}\mathbb{E}\left[Z^{2k}\right] = \mathbb{E}\left[\exp\left(2|t|Z^2\right) - 2|t|Z^2\right] .$$

Applying the work above, and using the fact $\frac{e}{\sqrt{\pi}} \leqslant 2$,

$$\leqslant 1 + 2|t|\mathbb{E}\left[Z^2\right] + \frac{32t^2C^2}{1 - 4|t|C} - 2|t|\mathbb{E}\left[Z^2\right] \leqslant \exp\left\{\frac{32t^2C^2}{1 - 4|t|C}\right\} .$$

□

D.8 Lower bounds on variance for signs of a Gaussian

Lemma D.4. *Take any positive definite $\Sigma \in \mathbb{R}^{p \times p}$, any distinct $a, b \in \{1, \ldots, p\}$, and any matrix $M \in \mathbb{R}^{p \times p}$ with $M_{ja} = 0$ for all j. Let $Z \sim N(0, \Sigma)$. Then there exists a constant C_{signs} depending on $\mathsf{C}(\Sigma)$ only, such that*

$$\mathsf{Var}\left(\mathrm{sign}(Z)^\top M \,\mathrm{sign}(Z)\right) \geqslant C_{\mathsf{signs}} \cdot M_{ab}^2 .$$

Proof of Lemma D.4. By the law of total variance,

$$\mathsf{Var}\left(\mathrm{sign}(Z)^\top M \,\mathrm{sign}(Z)\right) \geqslant \mathbb{E}\left[\mathsf{Var}\left(\mathrm{sign}(Z)^\top M \,\mathrm{sign}(Z) \mid Z_{(-a)}\right)\right] .$$

Let $(-a)$ denote the set $[p] \backslash \{a\}$. Let $M_{j,(-a)} \in \mathbb{R}^{p-1}$ denote the jth row of M with its ath entry removed, written as a column vector. Then, recalling that $M_{ja} = 0$ for all j, we have

$$\begin{aligned}&\mathsf{Var}\left(\mathrm{sign}(Z)^\top M \,\mathrm{sign}(Z) \mid Z_{(-a)}\right)\\&= \mathsf{Var}\left(\mathrm{sign}(Z_a) \cdot M_{a,(-a)}^\top \mathrm{sign}(Z_{(-a)}) + \sum_{j \neq a} \mathrm{sign}(Z_j) \cdot M_{j,(-a)}^\top \mathrm{sign}(Z_{(-a)}) \mid Z_{(-a)}\right) .\end{aligned}$$

Since every term except $\mathrm{sign}(Z_a)$ is a function of $Z_{(-a)}$,

$$= \mathsf{Var}\left(\mathrm{sign}(Z_a) \mid Z_{(-a)}\right) \cdot \left(M_{a,(-a)}^\top \mathrm{sign}(Z_{(-a)})\right)^2$$

Since the distribution of Z_a conditional on $Z_{(-a)}$ is given by $Z_{(-a)}^\top \beta_a + N(0, \nu_a^2)$ where $\beta_a = \Sigma_{(-a)}^{-1}\Sigma_{(-a),a}$ and $\nu_a^2 = \Sigma_{aa} - \Sigma_{(-a),a}^\top \Sigma_{(-a)}^{-1}\Sigma_{(-a),a}$,

$$\begin{aligned}&= \mathsf{Var}\left(\mathrm{sign}(Z_{(-a)}^\top \beta_a + N(0, \nu_a^2))\right) \cdot \left(M_{a,(-a)}^\top \mathrm{sign}(Z_{(-a)})\right)^2\\&= \left(1 - \mathbb{E}\left[\mathrm{sign}(Z_{(-a)}^\top \beta_a + N(0, \nu_a^2))\right]^2\right) \cdot \left(M_{a,(-a)}^\top \mathrm{sign}(Z_{(-a)})\right)^2\\&= \left(1 - \psi\left(\frac{Z_{(-a)}^\top \beta_a}{\nu_a}\right)^2\right) \cdot \left(M_{a,(-a)}^\top \mathrm{sign}(Z_{(-a)})\right)^2 ,\end{aligned}$$

where $\psi(x) = \Phi(x) - \Phi(-x) \in (-1, 1)$.

Now we will give a lower bound on the expectation of this quantity. First consider the term $\psi\left(\frac{Z_{(-a)}^\top \beta_a}{\nu_a}\right)$. Note that

$$\frac{Z_{(-a)}^\top \beta_a}{\nu_a} \sim N\left(0, \frac{\beta_a^\top \Sigma_{(-a)} \beta_a}{\nu_a^2}\right) = N\left(0, \frac{\Sigma_{(-a),a}^\top \Sigma_{(-a)}^{-1} \Sigma_{(-a),a}}{\Sigma_{aa} - \Sigma_{(-a),a}^\top \Sigma_{(-a)}^{-1} \Sigma_{(-a),a}}\right)$$

and this variance is bounded by $\mathsf{C}(\Sigma)$. Then, for any $c \in (0,1)$,

$$\mathbb{P}\left\{\left|\psi\left(\frac{Z_{(-a)}^\top \beta_a}{\nu_a}\right)\right| \leqslant \psi\left(\sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - c/2\right)\right)\right\} = \mathbb{P}\left\{\left|\frac{Z_{(-a)}^\top \beta_a}{\nu_a}\right| \leqslant \sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - c/2\right)\right\}$$
$$\geqslant \mathbb{P}\left\{|N(0, \mathsf{C}(\Sigma))| \leqslant \sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - c/2\right)\right\} = 1 - c\,. \quad \text{(D.9)}$$

Next, note that $M_{a,(-a)}^\top \operatorname{sign}(Z_{(-a)})$ is $\left(||M_{a,(-a)}||_2^2 \cdot \mathsf{C}(\Sigma)\right)$-subgaussian by Lemma 4.5, and

$$\mathbb{E}\left[\left(M_{a,(-a)}^\top \operatorname{sign}(Z_{(-a)})\right)^2\right] \geqslant ||M_{a,(-a)}||_2^2 \cdot \lambda_{\min}(T)\ ,$$

where $T = \mathbb{E}\left[\operatorname{sign}(Z)\operatorname{sign}(Z)^\top\right]$ (recall that $\Sigma = \sin\left(\frac{\pi}{2}T\right)$). Furthermore, by Wegkamp and Zhao (2013, Section 4.3), we have

$$T = \frac{2}{\pi}\sum_{k\geqslant 1} g(k) \Sigma \circ_k \Sigma\ ,$$

where $g(k) \geqslant 0$ are nonnegative scalars, $g(1) = 1$, and $\Sigma \circ_k \Sigma$ is the k-fold Hadamard product, that is, $(\Sigma \circ_k \Sigma)_{ij} = (\Sigma_{ij})^k$. Wegkamp and Zhao (2013, Section 4.3) show also that $\Sigma \circ_k \Sigma \succeq 0$ for all k. Therefore,

$$T = \frac{2}{\pi}\Sigma + \frac{2}{\pi}\sum_{k\geqslant 2} g(k)\Sigma \circ_k \Sigma \succeq \frac{2}{\pi}\Sigma\ ,$$

and so $\lambda_{\min}(T) \geqslant \frac{2}{\pi}\lambda_{\min}(\Sigma) \geqslant \frac{2}{\pi}(\mathsf{C}(\Sigma))^{-1}$. Applying Lemma D.6 (stated below),

$$\mathbb{P}\left\{\left(M_{a,(-a)}^\top \operatorname{sign}(Z_{(-a)})\right)^2 \geqslant ||M_{a,(-a)}||_2^2 \cdot \lambda_{\min}(T)/2\right\} \geqslant \frac{1}{16e^{2\mathsf{C}(\Sigma)/\lambda_{\min}(T)}}\ ,$$

and so,

$$\mathbb{P}\left\{\left(M_{a,(-a)}^\top \operatorname{sign}(Z_{(-a)})\right)^2 \geqslant ||M_{a,(-a)}||_2^2 \cdot \frac{1}{\pi\mathsf{C}(\Sigma)}\right\} \geqslant \frac{1}{16e^{\pi\mathsf{C}(\Sigma)^2}}\ .$$

Now set $c = \frac{1}{32e^{\pi\mathsf{C}(\Sigma)^2}}$ in (D.9). Then, we see that with probability at least $\frac{1}{32e^{\pi\mathsf{C}(\Sigma)^2}}$,

$$\left(1 - \psi\left(\frac{Z_{(-a)}^\top \beta_a}{\nu_a}\right)^2\right) \cdot \left(M_{a,(-a)}^\top \operatorname{sign}(Z_{(-a)})\right)^2 \geqslant$$
$$\left(1 - \psi\left(\sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - \frac{1}{64e^{\pi\mathsf{C}(\Sigma)^2}}\right)\right)^2\right) \cdot ||M_{a,(-a)}||_2^2 \cdot \frac{1}{\pi\mathsf{C}(\Sigma)}\ .$$

Therefore, combining everything,

$$\mathsf{Var}\left(\text{sign}(Z)^\top M\,\text{sign}(Z)\right) \geqslant \frac{1}{32e^{\pi\mathsf{C}(\Sigma)^2}} \cdot \left(1 - \psi\left(\sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - \frac{1}{64e^{\pi\mathsf{C}(\Sigma)^2}}\right)\right)^2\right) \cdot \frac{||M_{a,(-a)}||_2^2}{\pi\mathsf{C}(\Sigma)} \;.$$

Noting that $||M_{a,(-a)}||_2^2 \geqslant M_{ab}^2$, this proves the desired result, where we define

$$C_{\mathsf{signs}} = \frac{1}{32e^{\pi\mathsf{C}(\Sigma)^2}} \cdot \left(1 - \psi\left(\sqrt{\mathsf{C}(\Sigma)} \cdot \Phi^{-1}\left(1 - \frac{1}{64e^{\pi\mathsf{C}(\Sigma)^2}}\right)\right)^2\right) \cdot \frac{1}{\pi\mathsf{C}(\Sigma)} \;.$$

□

Lemma D.6. *Suppose that $W \in \mathbb{R}$ is a random variable with $\mathbb{E}[W] = 0$, $\mathbb{E}[W^2] \geqslant C_0$, and $\mathbb{E}[e^{tW}] \leqslant e^{C_1 t^2/2}$ for all $t \in \mathbb{R}$. Then*

$$\mathbb{P}\left\{W^2 \geqslant C_0/2\right\} \geqslant \frac{1}{16e^{2C_1/C_0}} \;.$$

Proof of Lemma D.6.

$$\begin{aligned} C_0/2 &\leqslant \mathbb{E}[W^2] - C_0/2 \\ &= \mathbb{E}[W^2 \cdot \mathbb{1}\left\{W^2 \geqslant C_0/2\right\}] + \mathbb{E}[W^2 \cdot \mathbb{1}\left\{W^2 < C_0/2\right\}] - C_0/2 \\ &\leqslant \mathbb{E}[W^2 \cdot \mathbb{1}\left\{W^2 \geqslant C_0/2\right\}] \;. \end{aligned}$$

Since $t^2 \leqslant e^t + e^{-t}$ for all $t \in \mathbb{R}$,

$$\begin{aligned} &\leqslant C_0\mathbb{E}[(e^{W/\sqrt{C_0}} + e^{-W/\sqrt{C_0}}) \cdot \mathbb{1}\left\{W^2 \geqslant C_0/2\right\}] \\ &= C_0\mathbb{E}[e^{W/\sqrt{C_0}} \cdot \mathbb{1}\left\{W^2 \geqslant C_0/2\right\}] + C_0\mathbb{E}[e^{-W/\sqrt{C_0}} \cdot \mathbb{1}\left\{W^2 \geqslant C_0/2\right\}] \\ &\leqslant C_0\sqrt{\mathbb{E}[(e^{W/\sqrt{C_0}})^2] \cdot \mathbb{E}[\mathbb{1}\left\{W^2 \geqslant C_0/2\right\}^2]} + C_0\sqrt{\mathbb{E}[(e^{-W/\sqrt{C_0}})^2] \cdot \mathbb{E}[\mathbb{1}\left\{W^2 \geqslant C_0/2\right\}^2]} \\ &= C_0\sqrt{\mathbb{E}[e^{2W/\sqrt{C_0}}] \cdot \mathbb{P}\{W^2 \geqslant C_0/2\}} + C_0\sqrt{\mathbb{E}[e^{-2W/\sqrt{C_0}}] \cdot \mathbb{P}\{W^2 \geqslant C_0/2\}} \\ &\leqslant C_0\sqrt{e^{C_1/C_0 \cdot 2^2/2} \cdot \mathbb{P}\{W^2 \geqslant C_0/2\}} + C_0\sqrt{e^{C_1/C_0 \cdot 2^2/2} \cdot \mathbb{P}\{W^2 \geqslant C_0/2\}} \;, \end{aligned}$$

and rearranging terms we have proved the lemma. □

D.9 Bounding the error in estimating the variance (Lemma B.5)

Lemma B.5. *Under the assumptions and definitions of Theorem 4.2, with probability at least $1 - \frac{1}{6p_n}$, if $n \geqslant k_n^2 \log(p_n)$, on the event that the bounds (3.1) in Assumption 3.3 hold,*

$$\left|\check{S}_{ab} \cdot \det(\check{\Theta}) - S_{ab} \cdot \det(\Theta)\right| \leqslant C_{\mathsf{oracle}} \cdot \sqrt{\frac{k_n^2 \log(p_n)}{n}} \;.$$

Proof of Lemma B.5. Recall from the proof of Theorem 4.1 that we have defined

$$g(X,X') = \mathrm{sign}(X-X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \mathrm{sign}(X-X')\ ,$$

and $g_1(X) = \mathbb{E}\left[g(X,X') \mid X\right]$, where

$$u_a = 1, u_b = 0, u_I = -\gamma_a \text{ and } v_a = 0, v_b = 1, v_I = -\gamma_b\ .$$

Recall from the proof of Lemma B.1, given in Appendix D.6, that we have

$$\nu_{g_1}^2 = \mathsf{Var}(g_1(X)) = \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)^\top \cdot \Sigma_{h_1} \cdot \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)\ ,$$

where $\Sigma_{h_1} = \mathsf{Var}(h_1(X))$ for

$$h_1(X) = \mathbb{E}\left[\mathrm{sign}(X-X') \otimes \mathrm{sign}(X-X') \mid X\right] \in \mathbb{R}^{p_n^2}\ .$$

To estimate this variance, define vectors $\breve{u}$ and $\breve{v}$ with entries

$$\breve{u}_a = 1, \breve{u}_b = 0, \breve{u}_I = -\breve{\gamma}_a \text{ and } \breve{v}_a = 0, \breve{v}_b = 1, \breve{v}_I = -\breve{\gamma}_b\ ,$$

and define

$$\widehat{\Sigma}_{h_1} = \frac{1}{n}\sum_i \left(\widehat{h}_1(X_i) - \frac{1}{n}\sum_{i'} \widehat{h}_1(X_{i'})\right)\left(\widehat{h}_1(X_i) - \frac{1}{n}\sum_{i'} \widehat{h}_1(X_{i'})\right)^\top\ ,$$

where abusing notation, we write

$$\widehat{h}_1(X_i) = \frac{1}{n-1}\sum_{i'\neq i} h(X_i, X_i') = \frac{1}{n-1}\sum_{i'\neq i} \mathrm{sign}(X_i - X_{i'}) \otimes \mathrm{sign}(X_i - X_{i'})\ .$$

We then define

$$\breve{\nu}_{g_1}^2 = \mathsf{vec}\left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}\widehat{T}\right)\right)^\top \cdot \widehat{\Sigma}_{h_1} \cdot \mathsf{vec}\left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}\widehat{T}\right)\right)\ .$$

Writing

$$x = \mathsf{vec}\left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \text{ and } \breve{x} = \mathsf{vec}\left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}\widehat{T}\right)\right)\ ,$$

we then have

$$\nu_{g_1}^2 = x^\top \Sigma_{h_1} x \text{ and } \breve{\nu}_{g_1}^2 = \breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}\ .$$

Define also

$$\overline{x} = \mathsf{vec}\left(\breve{u}\breve{v}^\top \circ \cos\left(\frac{\pi}{2}T\right)\right)\ .$$

The following lemma, proved in Appendix D.10, carries out some elementary calculations on the norms of these vectors $x, \overline{x}, \breve{x}$.

Lemma D.7. *Define $x, \overline{x}, \breve{x}$ as in the proof of Lemma B.5, and assume $n \geqslant k_n^2 \log(p_n)$. With probability at least $1-\frac{1}{36p_n}$, if the bounds (3.1) in Assumption 3.3 hold then the following inequalities all hold for constants C_0, C_1, C_2, C_3 that depend only on $C_{\mathsf{cov}}, C_{\mathsf{sparse}}, C_{\mathsf{est}}$:*

$$\begin{aligned} ||x||_1 &\leqslant C_0 k_n \ , \\ ||\breve{x} - \overline{x}||_1 &\leqslant C_1 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ , \\ ||\breve{x} - x||_1 &\leqslant C_2 \sqrt{\frac{k_n^3 \log(p_n)}{n}} \ , \\ ||\mathsf{mat}(\overline{x} - x)||_{\ell_1/\ell_2} &\leqslant C_3 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ , \end{aligned}$$

where $\mathsf{mat}(\cdot)$ reshapes a vector in $\mathbb{R}^{p_n^2}$ into a $p_n \times p_n$ matrix, and where we define the matrix ℓ_1/ℓ_2 norm as $M_{\ell_1/\ell_2} := \sum_j ||M_j||_2$, where M_j is the jth column of M.

We now continue bounding error in estimating ν_{g_1}. We have:

$$\left|\breve{\nu}_{g_1}^2 - \nu_{g_1}^2\right| = \left|\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x} - x^\top \Sigma_{h_1} x\right| \leqslant \left|x^\top (\widehat{\Sigma}_{h_1} - \Sigma_{h_1}) x\right| + \left|\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x} - x^\top \widehat{\Sigma}_{h_1} x\right| \ . \tag{D.10}$$

We bound each term separately. For the first term in (D.10), we apply the following lemma (proved in Appendix D.10 below):

Lemma D.8. *Under the same assumptions and notation as Lemmas B.1 and B.5, for a universal constant $C_{\mathsf{studentized}}$,*

$$\mathbb{P}\left\{\left|x^\top (\widehat{\Sigma}_{h_1} - \Sigma_{h_1}) x\right| \leqslant C_{\mathsf{studentized}} \sqrt{\frac{k_n^2 \log(p_n)}{n}}\right\} \geqslant 1 - \frac{1}{36 p_n} \ .$$

For the second term in (D.10), since $\widehat{\Sigma}_{h_1} \succeq 0$ and so $y \mapsto \sqrt{y^\top \widehat{\Sigma}_{h_1} y}$ is a norm and must satisfy the triangle inequality,

$$\begin{aligned} \left|\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x} - x^\top \widehat{\Sigma}_{h_1} x\right| &= \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \cdot \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} + \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \\ &\leqslant \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right|^2 + \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \cdot 2\sqrt{x^\top \widehat{\Sigma}_{h_1} x} \\ &\leqslant \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right|^2 + \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \cdot 2\sqrt{x^\top \Sigma_{h_1} x + \left|x^\top (\widehat{\Sigma}_{h_1} - \Sigma_{h_1}) x\right|} \ . \end{aligned} \tag{D.11}$$

To bound the difference term $\left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right|$ which appears twice in the expression above, we have

$$\begin{aligned}\left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| &\leqslant \sqrt{(\breve{x}-x)^\top \widehat{\Sigma}_{h_1} (\breve{x}-x)} \\ &\leqslant \sqrt{(\breve{x}-x)^\top \Sigma_{h_1} (\breve{x}-x)} + \sqrt{\left|(\breve{x}-x)^\top (\widehat{\Sigma}_{h_1} - \Sigma_{h_1})(\breve{x}-x)\right|} \\ &\leqslant \sqrt{(\breve{x}-x)^\top \Sigma_{h_1} (\breve{x}-x)} + \sqrt{||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty \cdot ||\breve{x}-x||_1^2} \\ &\leqslant \sqrt{(\breve{x}-\overline{x})^\top \Sigma_{h_1} (\breve{x}-\overline{x})} + \sqrt{(\overline{x}-x)^\top \Sigma_{h_1} (\overline{x}-x)} + \sqrt{||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty \cdot ||\breve{x}-x||_1^2} \\ &\leqslant \sqrt{||\Sigma_{h_1}||_\infty ||\breve{x}-\overline{x}||_1^2} + \sqrt{(\overline{x}-x)^\top \Sigma_{h_1} (\overline{x}-x)} + \sqrt{||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty \cdot ||\breve{x}-x||_1^2} \\ &\leqslant C_1 \sqrt{\frac{k_n^2 \log(p_n)}{n}} + \sqrt{(\overline{x}-x)^\top \Sigma_{h_1} (\overline{x}-x)} + \sqrt{||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty} \cdot C_2 \sqrt{\frac{k_n^3 \log(p_n)}{n}} \ . \qquad \text{(D.12)}\end{aligned}$$

Next, we state two lemmas, which are proved in Appendix D.10.

Lemma D.9. *With probability at least* $1 - \frac{1}{9p_n}$,

$$||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty \leqslant 100\sqrt{\frac{\log(p_n)}{n}} \ .$$

Lemma D.10. *Let* Σ_{h_1} *be defined as in Assumption 3.4. For every* $z \in \mathbb{R}^{p_n^2}$,

$$z^\top \Sigma_{h_1} z \leqslant \lambda_{\max}(\Sigma) \cdot ||\mathsf{mat}(z)||_{\ell_1/\ell_2}^2 \ ,$$

where $||\mathsf{mat}(z)||_{\ell_1/\ell_2}$ *is defined as in the statement of Lemma D.7.*

From this point on, we assume that the bounds derived in Lemmas D.7 and D.9 all hold (which the lemmas have shown to be true with probability at least $1 - \frac{1}{6p_n}$, on the event that the bounds (3.1) of Assumption 3.3 hold.) By Lemmas D.10 and D.7,

$$(\overline{x}-x)^\top \Sigma_{h_1} (\overline{x}-x) \leqslant C_{\mathsf{cov}} \cdot \left(C_3\sqrt{\frac{k_n^2 \log(p_n)}{n}}\right)^2 \ .$$

Applying this bound, along with the high probability events of Lemmas D.8 and D.9, we return to (D.12) and obtain

$$\begin{aligned}&\left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \leqslant \\ &\quad C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + \sqrt{C_{\mathsf{cov}} \cdot \left(C_3\sqrt{\frac{k_n^2 \log(p_n)}{n}}\right)^2} + \sqrt{100\sqrt{\frac{\log(p_n)}{n}}} \cdot C_2\sqrt{\frac{k_n^3 \log(p_n)}{n}} \\ &\qquad = \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot \left(C_1 + C_3\sqrt{C_{\mathsf{cov}}} + 10C_2\sqrt[4]{\frac{k_n^2 \log(p_n)}{n}}\right) \leqslant \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot C_4 \ ,\end{aligned}$$

where for the last step we define $C_4 = C_1 + C_3\sqrt{C_{\mathsf{cov}}} + 10C_2$ and use the assumption $n \geqslant k_n^2 \log(p_n)$. Next, returning to (D.11),

$$\begin{aligned}
&\left|\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x} - x^\top \widehat{\Sigma}_{h_1} x\right| \\
&\quad \leqslant \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right|^2 + \left|\sqrt{\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x}} - \sqrt{x^\top \widehat{\Sigma}_{h_1} x}\right| \cdot 2\sqrt{x^\top \Sigma_{h_1} x + \left|x^\top (\widehat{\Sigma}_{h_1} - \Sigma_{h_1}) x\right|} \\
&\qquad\qquad \leqslant C_4^2 \cdot \frac{k_n^2 \log(p_n)}{n} + C_4 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot 2\sqrt{C_{\mathsf{moment}}^{2/3} + C_{\mathsf{studentized}} \sqrt{\frac{k_n^2 \log(p_n)}{n}}} \ ,
\end{aligned}$$

where the last step applies the high probability event of Lemma D.8, and uses the fact that $x^\top \Sigma_{h_1} x = \nu_{g_1}^2 \leqslant C_{\mathsf{moment}}^{2/3}$ by Lemma B.1. Defining $C_5 = C_4^2 + C_4 \cdot 2\sqrt{C_{\mathsf{moment}}^{2/3} + C_{\mathsf{studentized}}}$, and using the assumption $n \geqslant k^2 \log(p_n)$, we have

$$\left|\breve{x}^\top \widehat{\Sigma}_{h_1} \breve{x} - x^\top \widehat{\Sigma}_{h_1} x\right| \leqslant C_5 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ .$$

Finally, returning to (D.10) and applying Lemma D.9, we see that

$$|\breve{\nu}_{g_1}^2 - \nu_{g_1}^2| \leqslant C_{\mathsf{studentized}} \cdot \sqrt{\frac{k_n^2 \log(p_n)}{n}} + C_5 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ .$$

Next, we have

$$|\breve{\nu}_{g_1} - \nu_{g_1}| = \frac{\left|\breve{\nu}_{g_1}^2 - \nu_{g_1}^2\right|}{\breve{\nu}_{g_1} + \nu_{g_1}} \leqslant \frac{\left|\breve{\nu}_{g_1}^2 - \nu_{g_1}^2\right|}{\nu_{g_1}} \leqslant \frac{(C_{\mathsf{studentized}} + C_5)\sqrt{\frac{k_n^2 \log(p_n)}{n}}}{\frac{1}{\pi} C_{\mathsf{variance}}} \ ,$$

where for the denominator we apply Lemma B.1. Finally, since we know that $S_{ab} = \pi \nu_{g_1} \cdot (\det(\Theta))^{-1}$ and $\breve{S}_{ab} = \pi \breve{\nu}_{g_1} \cdot (\det(\breve{\Theta}))^{-1}$, and then we have

$$\left|\breve{S}_{ab} \cdot \det(\breve{\Theta}) - S_{ab} \cdot \det(\Theta)\right| = \pi \cdot |\breve{\nu}_{g_1} - \nu_{g_1}| \leqslant \pi \cdot \frac{(C_{\mathsf{studentized}} + C_5)\sqrt{\frac{k_n^2 \log(p_n)}{n}}}{\frac{1}{\pi} C_{\mathsf{variance}}} \ .$$

Defining

$$C_{\mathsf{oracle}} \geqslant \pi \cdot \frac{C_{\mathsf{studentized}} + C_5}{\frac{1}{\pi} C_{\mathsf{variance}}} \ ,$$

we see that

$$\left|\breve{S}_{ab} \cdot \det(\breve{\Theta}) - S_{ab} \cdot \det(\Theta)\right| \leqslant C_{\mathsf{oracle}} \cdot \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ .$$

□

D.10 Calculations for the variance estimate (Lemma B.5)

Proof of Lemma D.7. We calculate

$$||x||_1 = ||uv^\top \circ \cos\left(\frac{\pi}{2} T\right)||_1 \leqslant ||uv^\top||_1 \cdot ||\cos\left(\frac{\pi}{2} T\right)||_\infty \leqslant ||u||_1 ||v||_1 \leqslant k_n (1 + 2C_{\mathsf{cov}} C_{\mathsf{sparse}})^2 =: C_0 k_n \ ,$$

where for the last inequality we apply (B.2).

Next,

$$
\begin{aligned}
||\breve{x} - \overline{x}||_1 &= ||\breve{u}\breve{v}^\top \circ \left(\cos\left(\frac{\pi}{2}T\right) - \cos\left(\frac{\pi}{2}\widehat{T}\right)\right)||_1 \\
&\leqslant (||u||_1 + ||\breve{u} - u||_1) \cdot (||v||_1 + ||\breve{v} - v||_1) \cdot || \cos\left(\frac{\pi}{2}T\right) - \cos\left(\frac{\pi}{2}\widehat{T}\right)||_\infty
\end{aligned}
$$

Applying (B.2) and Assumption 3.3, and the fact that $\cos(\cdot)$ is 1-Lipschitz, if the bounds in Assumption 3.3 hold,

$$
\leqslant \left(\sqrt{k_n}(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) + C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}}\right)^2 \cdot \frac{\pi}{2}||\widehat{T} - T||_\infty
$$

Applying Lemma B.2, with probability at least $1 - \frac{1}{36p_n}$,

$$
\leqslant \left(\sqrt{k_n}(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) + C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}}\right)^2 \cdot \frac{\pi}{2}\sqrt{\frac{12\log(36p_n)}{n}}
$$

Since $12\log(36p_n) \leqslant 108\log(p_n) \leqslant 4n$, where the last step holds by assumption in Theorem 4.2,

$$
\begin{aligned}
&\leqslant \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot ((1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) + C_{\mathsf{est}})^2 \cdot \pi \\
&= C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} \text{ for } C_1 := ((1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) + C_{\mathsf{est}})^2 \cdot \pi \ .
\end{aligned}
$$

Next,

$$
\begin{aligned}
||\breve{x} - x||_1 &\leqslant ||\breve{x} - \overline{x}||_1 + ||\overline{x} - x||_1 \\
&\leqslant C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + ||\overline{x} - x||_1 \\
&= C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + ||(\breve{u}\breve{v}^\top - uv^\top) \circ \cos\left(\frac{\pi}{2}T\right)||_1 \\
&\leqslant C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + ||\breve{u}(\breve{v} - v)^\top \circ \cos\left(\frac{\pi}{2}T\right)||_1 + ||(\breve{u} - u)v^\top \circ \cos\left(\frac{\pi}{2}T\right)||_1 \\
&\leqslant C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + ||\breve{u}||_1||\breve{v} - v||_1|| \cos\left(\frac{\pi}{2}T\right)||_\infty + ||\breve{u} - u||_1||v||_1|| \cos\left(\frac{\pi}{2}T\right)||_\infty
\end{aligned}
$$

Applying (B.2), if the bounds (3.1) in Assumption 3.3 hold,

$$
\begin{aligned}
&\leqslant C_1\sqrt{\frac{k_n^2 \log(p_n)}{n}} + \left(\sqrt{k_n}(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) + C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}}\right) \cdot C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}} \\
&\qquad\qquad + \sqrt{k_n}(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) \cdot C_{\mathsf{est}}\sqrt{\frac{k_n^2 \log(p_n)}{n}} \\
&\leqslant C_2\sqrt{\frac{k_n^3 \log(p_n)}{n}} \ ,
\end{aligned}
$$

where $C_2 = C_1 + 2(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}}) \cdot C_{\mathsf{est}} + C_{\mathsf{est}}^2$ and we use the assumption $n \geqslant k_n^2 \log(p)$.

Finally, noting that $\overline{x} - x = \mathsf{vec}\left((\breve{u}\breve{v}^\top - uv^\top) \circ \cos\left(\frac{\pi}{2}T\right)\right)$, we calculate the ℓ_1/ℓ_2 norm of this matrix:

$$\begin{aligned}
&||\mathsf{mat}(\overline{x} - x)||_{\ell_1/\ell_2} = ||(\breve{u}\breve{v}^\top - uv^\top) \circ \cos\left(\frac{\pi}{2}T\right)||_{\ell_1/\ell_2} \\
&= \sum_j || \left[(\breve{u}\breve{v}^\top - uv^\top) \circ \cos\left(\frac{\pi}{2}T\right)\right]_j ||_2 \\
&\leqslant \sum_j || \left[\breve{u}\breve{v}^\top - uv^\top\right]_j ||_2 \cdot || \cos\left(\frac{\pi}{2}T\right) ||_\infty \\
&\leqslant \sum_j ||\breve{u} \cdot (\breve{v}_j - v_j)||_2 + ||(\breve{u} - u) \cdot v_j||_2 \\
&= \sum_j ||\breve{u}||_2 \cdot |\breve{v}_j - v_j| + ||\breve{u} - u||_2 \cdot |v_j| \\
&= ||\breve{u}||_2 ||\breve{v} - v||_1 + ||\breve{u} - u||_2 ||v||_1 \ .
\end{aligned}$$

Applying (B.1), if the bounds (3.1) in Assumption 3.3 hold,

$$\begin{aligned}
&\leqslant \left(\sqrt{1 + C_{\mathsf{cov}}^2} + C_{\mathsf{est}}\sqrt{\frac{k_n \log(p_n)}{n}}\right) \cdot C_{\mathsf{est}} \cdot \sqrt{\frac{k_n^2 \log(p_n)}{n}} + C_{\mathsf{est}} \cdot \sqrt{\frac{k_n \log(p_n)}{n}} \cdot \sqrt{k_n}\sqrt{1 + C_{\mathsf{cov}}} \\
&\leqslant C_3 \sqrt{\frac{k_n^2 \log(p_n)}{n}} \ ,
\end{aligned}$$

where we define $C_3 = 2\sqrt{1 + C_{\mathsf{cov}}^2} \cdot C_{\mathsf{est}} + C_{\mathsf{est}}^2$ and use the assumption that $n \geqslant k_n^2 \log(p_n)$. □

Proof of Lemma D.8. By definition, we have $\Sigma_{h_1} = \mathsf{Var}(h_1(X))$ for

$$h_1(X) = \mathbb{E}\left[\text{sign}(X - X') \otimes \text{sign}(X - X') \mid X\right] \in \mathbb{R}^{p_n^2} \ .$$

Therefore, since x is fixed,

$$x^\top \Sigma_{h_1} x = x^\top \mathsf{Var}(h_1(X)) x = \mathsf{Var}(x^\top h_1(X)) = \mathsf{Var}(g_1(X)) = \nu_{g_1}^2 \ ,$$

where we recall that $g_1(X) = \mathbb{E}\left[g(X, X') \mid X\right]$ where we define the kernel

$$g(X, X') = \text{sign}(X - X')^\top \left(uv^\top \circ \cos\left(\frac{\pi}{2}T\right)\right) \text{sign}(X - X') = x^\top h(X, X') \ .$$

Define

$$\gamma(X, X', X'') = \frac{g(X, X')g(X, X'') + g(X', X)g(X', X'') + g(X'', X)g(X'', X')}{3} \ .$$

Note that γ is a U-statistic of order 3, with

$$\begin{aligned}
||\gamma||_\infty := \sup_{X, X', X''} |\gamma(X, X', X'')| \leqslant \sup_{X, X'} |g(X, X')|^2 \leqslant ||x||_1^2 \sup_{X, X'} ||h(X, X')||_\infty^2 \\
= ||x||_1^2 \leqslant k_n^2 \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4 \ .
\end{aligned}$$

(See proof of Lemma D.7 for this bound on $||x||_1$.) And,

$$\begin{aligned}\mathsf{Var}(\gamma) := \mathsf{Var}(\gamma(X,X',X'')) \leqslant \mathsf{Var}(g(X,X')g(X,X'')) \leqslant \mathbb{E}\left[|g(X,X')|^4\right] \\ \leqslant \mathbb{E}\left[|g(X,X')|^3\right] \cdot k_n \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2 \leqslant C_{\mathsf{moment}} \cdot k_n \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2 ,\end{aligned}$$

where we use Lemma B.1 for the last bound.

Next, we have

$$\begin{aligned}\mathbb{E}\left[\gamma(X,X',X'')\right] = \mathbb{E}\left[g(X,X')g(X,X'')\right] = \mathbb{E}\left[\mathbb{E}\left[g(X,X')g(X,X'')|X\right]\right] \\ = \mathbb{E}\left[\mathbb{E}\left[g(X,X')|X\right]\mathbb{E}\left[g(X,X'')|X\right]\right] = \mathbb{E}\left[g_1(X)^2\right] .\end{aligned}$$

Therefore,

$$x^\top \Sigma_{h_1} x = \nu_{g_1}^2 = \mathsf{Var}(g_1(X)) = \mathbb{E}\left[\gamma(X,X',X'')\right] - \mathbb{E}\left[g_1(X)\right]^2 = \mathbb{E}\left[\gamma(X,X',X'')\right] - \mathbb{E}\left[g(X,X')\right]^2 .$$

Next, examining the definition of $\widehat{\Sigma}_{h_1}$, we obtain

$$x^\top \widehat{\Sigma}_{h_1} x = \frac{1}{n(n-1)^2}\left[\sum_{i\neq i'\neq i''} \gamma(X_i,X_{i'},X_{i''}) + \sum_{i\neq i'} g(X_i,X_{i'})^2\right] - \left(\frac{1}{\binom{n}{2}}\sum_{i<i'} g(X_i,X_{i'})\right)^2 .$$

Therefore, using the fact that $|g(X,X')| \leqslant k_n \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2$ always,

$$\begin{aligned}\left|x^\top \widehat{\Sigma}_{h_1} x - x^\top \Sigma_{h_1} x\right| \leqslant \left|\frac{1}{\binom{n}{3}}\sum_{i<i'<i''} \gamma(X_i,X_{i'},X_{i''}) - \mathbb{E}[\gamma(X,X',X'')]\right| \\ + \frac{k_n^2 \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4}{n-1} + \left|\left(\frac{1}{\binom{n}{2}}\sum_{i<i'} g(X_i,X_{i'})\right)^2 - \mathbb{E}[g(X,X')]^2\right| .\end{aligned}$$

Now, using Bernstein's inequality for U-statistics (Peel et al. (2010, Theorem 2)), for any $\delta > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{\binom{n}{3}}\sum_{i<i'<i''} \gamma(X_i,X_{i'},X_{i''}) - \mathbb{E}[\gamma(X,X',X'')]\right| > \sqrt{\frac{2\mathsf{Var}(\gamma)\log(2/\delta)}{(n/3)}} + \frac{2||\gamma||_\infty \log(2/\delta)}{3(n/3)}\right\} \leqslant \delta .$$

Therefore, with probability at least $1 - \frac{1}{72p_n}$,

$$\begin{aligned}&\left|\frac{1}{\binom{n}{3}}\sum_{i<i'<i''} \gamma(X_i,X_{i'},X_{i''}) - \mathbb{E}[\gamma(X,X',X'')]\right| \leqslant \\ &\sqrt{\frac{2C_{\mathsf{moment}} \cdot k_n \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2 \log(2\cdot 72p_n)}{(n/3)}} + \frac{2k_n^2 \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4 \log(2\cdot 72p_n)}{3(n/3)} \\ &\leqslant \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot C' ,\end{aligned}$$

where

$$C' = \sqrt{6C_{\mathsf{moment}} \cdot (1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2(2+\log_2(72))} + 2\cdot(1+2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4(2+\log_2(72)) ,$$

and we use the assumption $n \geqslant k_n^2 \log(p_n)$ and $p_n \geqslant 2$. And, again using Bernstein's inequality for U-statistics, and using the fact that $|g(X, X')| \leqslant k_n \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2$ always, with probability at least $1 - \frac{1}{72p_n}$,

$$\left|\frac{1}{\binom{n}{2}} \sum_{i<i'} g(X_i, X_{i'}) - \mathbb{E}[g(X, X')]\right| \leqslant \\ \sqrt{\frac{2k_n^2 \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4 \log(2 \cdot 72p_n)}{(n/2)} + \frac{2k_n \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2 \log(2 \cdot 72p_n)}{3(n/2)}} \\ \leqslant \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot C'' ,$$

where

$$C'' = \sqrt{\frac{2 \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4 (2 + \log_2(72))}{(1/2)} + \frac{2 \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^2 (2 + \log_2(72))}{3/2}} ,$$

and we use the assumption $n \geqslant k_n^2 \log(p_n)$ and $p_n \geqslant 2$. Therefore,

$$\left|\left(\frac{1}{\binom{n}{2}} \sum_{i<i'} g(X_i, X_{i'})\right)^2 - \mathbb{E}[g(X, X')]^2\right| \leqslant \left|\frac{1}{\binom{n}{2}} \sum_{i<i'} g(X_i, X_{i'}) - \mathbb{E}[g(X, X')]\right|^2 + \\ \left|\frac{1}{\binom{n}{2}} \sum_{i<i'} g(X_i, X_{i'}) - \mathbb{E}[g(X, X')]\right| \cdot 2|\mathbb{E}[g(X, X')]| \leqslant C''' \sqrt{\frac{k_n^2 \log(p_n)}{n}} ,$$

where we set

$$C''' = C''^2 + 2C'' \cdot C_{\mathsf{moment}}^{1/3}$$

and again use $n \geqslant k_n^2 \log(p_n)$, and apply Lemma B.1 to bound $|\mathbb{E}\left[g(X, X')\right]|$. Combining everything, this proves that, with probability at least $1 - \frac{1}{36p_n}$,

$$\left|x^\top \widehat{\Sigma}_{h_1} x - x^\top \Sigma_{h_1} x\right| \leqslant \sqrt{\frac{k_n^2 \log(p_n)}{n}} \cdot C' + \frac{k_n^2 \cdot (1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4}{n-1} + C''' \sqrt{\frac{k_n^2 \log(p_n)}{n}} .$$

Setting

$$C_{\mathsf{studentized}} = C' + C''' + 2(1 + 2C_{\mathsf{cov}}C_{\mathsf{sparse}})^4$$

and using the fact that $n \geqslant 2$ and $n \geqslant k_n^2 \log(p_n)$, we have

$$\left|x^\top \widehat{\Sigma}_{h_1} x - x^\top \Sigma_{h_1} x\right| \leqslant C_{\mathsf{studentized}} \sqrt{\frac{k_n^2 \log(p_n)}{n}} .$$

□

Proof of Lemma D.9. From our definitions, we see that

$$\Sigma_{h_1} = \mathsf{Var}(h_1(X)) = \mathbb{E}\left[h_1(X) h_1(X)^\top\right] - \mathbb{E}[h_1(X)]\mathbb{E}[h_1(X)]^\top ,$$

and

$$\widehat{\Sigma}_{h_1} = \frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)^\top .$$

First, we bound $||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \mathbb{E}\left[h_1(X)h_1(X)^\top\right]||_\infty$. We have

$$\begin{aligned}
||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \mathbb{E}\left[h_1(X)h_1(X)^\top\right]||_\infty \leqslant& \\
&||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top||_\infty + \\
&\qquad ||\frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top - \mathbb{E}\left[h_1(X)h_1(X)^\top\right]||_\infty . \qquad \text{(D.13)}
\end{aligned}$$

We handle these two terms separately. First, we bound $||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top||_\infty$. For convenience we define $A := \frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top$ and $B := \frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top$. Since A and B are both positive semidefinite matrices with ones on the diagonal, we have

$$||A - B||_\infty = \frac{1}{2}\max_{j,k\in[p_n^2]}\left|f_{jk}^\top(A-B)f_{jk}\right| , \qquad \text{(D.14)}$$

where $f_{jk} \in \mathbb{R}^{p_n^2}$ is the vector with $(f_{jk})_j = 1$, $(f_{jk})_k = -1$, and zeros elsewhere. Next we have

$$\begin{aligned}
\left|f_{jk}^\top(A-B)f_{jk}\right| &= \left|\sqrt{f_{jk}^\top A f_{jk}} - \sqrt{f_{jk}^\top B f_{jk}}\right| \cdot \left(\sqrt{f_{jk}^\top A f_{jk}} + \sqrt{f_{jk}^\top B f_{jk}}\right) \\
&\leqslant 4\left|\sqrt{f_{jk}^\top A f_{jk}} - \sqrt{f_{jk}^\top B f_{jk}}\right| = \frac{4}{\sqrt{n}}\left|\sqrt{\sum_i (\widehat{h}_1(X_i)^\top f_{jk})^2} - \sqrt{\sum_i (h_1(X_i)^\top f_{jk})^2}\right| \\
&\leqslant \frac{4}{\sqrt{n}}\sqrt{\sum_i \left((\widehat{h}_1(X_i) - h_1(X_i))^\top f_{jk}\right)^2} ,
\end{aligned}$$

where the first inequality follows from the fact that $||f_{jk}||_1 \leqslant 2$ while $||A||_\infty, ||B||_\infty \leqslant 1$, and the second inequality follows from the triangle inequality. Next, for each i and each j, k, observe that

$$\widehat{h}_1(X_i)^\top f_{jk} = \frac{1}{n-1}\sum_{i'\neq i}\left(\text{sign}(X_i - X_{i'}) \otimes \text{sign}(X_i - X_{i'})\right)^\top f_{jk} ,$$

which after conditioning on X_i, is a mean of $(n-1)$ i.i.d. variables, each taking values in $[-2, 2]$ since $||f_{jk}||_1 \leqslant 2$. Furthermore, conditioning on X_i, we have $\mathbb{E}[\widehat{h}_1(X_i)] = h_1(X_i)$. Therefore, applying Hoeffding's lemma (see, for example, Lemma 2.6 in Massart, 2007), for each i, j, k, for any $t \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{t\cdot(\widehat{h}_1(X_i) - h_1(X_i))^\top f_{jk}\right\}\right] \leqslant \exp\left\{\frac{2t^2}{n-1}\right\} . \qquad \text{(D.15)}$$

Applying Lemma D.11 (stated below), then,

$$\mathbb{P}\left\{\frac{1}{n}\sum_i\left((\widehat{h}_1(X_i) - h_1(X_i))^\top f_{jk}\right)^2 > \frac{80}{n-1}\cdot(1 + \log(27p_n^5))\right\} \leqslant \frac{1}{27p_n^5} .$$

Taking a union bound over all $j, k \in [p_n^2]$, and returning to (D.14), we then have

$$\mathbb{P}\left\{||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top||_\infty > 2\sqrt{\frac{80}{n-1}\cdot(1+\log(27p_n^5))}\right\} \leqslant \frac{1}{27p_n} \ .$$

Next we turn to the second term in (D.13). Since $||h_1(X)||_\infty \leqslant 1$ always, we see that for each $j, k \in [p_n]$,

$$\left(\frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top\right)_{jk}$$

is a mean of n i.i.d. terms, each taking values in $[-1, 1]$. Applying Hoeffding's inequality, for each j, k,

$$\mathbb{P}\left\{\left|\left(\frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top - \mathbb{E}[h(X)h(X)^\top]\right)_{jk}\right| \geqslant t\right\} \leqslant 2e^{-nt^2/2}$$

for any $t \geqslant 0$. Setting $t = \sqrt{\frac{2\log(54p_n^3)}{n}}$, and taking a union bound, we see that

$$\mathbb{P}\left\{||\frac{1}{n}\sum_i h_1(X_i)h_1(X_i)^\top - \mathbb{E}[h(X)h(X)^\top]||_\infty \geqslant \sqrt{\frac{2\log(54p_n^3)}{n}}\right\} \leqslant 2p_n^2{\cdot}e^{-n\left(\sqrt{\frac{54\log(p_n^3)}{n}}\right)^2/2} = \frac{1}{27p_n} \ .$$

Returning to (D.13), then, with probability at least $1 - \frac{2}{27p_n}$,

$$||\frac{1}{n}\sum_i \widehat{h}_1(X_i)\widehat{h}_1(X_i)^\top - \mathbb{E}\left[h_1(X)h_1(X)^\top\right]||_\infty \leqslant 2\sqrt{\frac{80}{n-1}\cdot(1+\log(27p_n^3))} + \sqrt{\frac{2\log(54p_n^3)}{n}} \ . \tag{D.16}$$

Next, to complete the proof, we bound

$$||\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)^\top - \mathbb{E}\left[h_1(X)\right]\mathbb{E}\left[h_1(X)\right]^\top||_\infty \ .$$

We have

$$\begin{aligned}&\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)^\top - \mathbb{E}\left[h_1(X)\right]\mathbb{E}\left[h_1(X)\right]^\top \\ &\quad = \left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]\right)^\top - \left(\frac{1}{n}\sum_i \widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]\right)\mathbb{E}\left[h_1(X)\right]^\top\end{aligned}$$

and, since $||\frac{1}{n}\sum_i \widehat{h}_1(X_i)||_\infty, ||\mathbb{E}\left[h_1(X)\right]||_\infty \leqslant 1$, we therefore have

$$||\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)\left(\frac{1}{n}\sum_i \widehat{h}_1(X_i)\right)^\top - \mathbb{E}\left[h_1(X)\right]\mathbb{E}\left[h_1(X)\right]^\top||_\infty \leqslant 2||\frac{1}{n}\sum_i \widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]||_\infty \ .$$

For each sign $s \in \{\pm 1\}$, for each $j \in [p_n^2]$, writing $\mathbf{e}_j$ to denote the jth basis vector in $\mathbb{R}^{p_n^2}$, we have

$$\mathbb{E}\left[\exp\left\{t \cdot s \cdot \mathbf{e}_j^\top \left(\frac{1}{n}\sum_i \widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]\right)\right\}\right] \leqslant \frac{1}{n}\sum_i \mathbb{E}\left[\exp\left\{t \cdot s \cdot \mathbf{e}_j^\top \left(\widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]\right)\right\}\right]$$
$$\leqslant \exp\left\{\frac{t^2}{2(n-1)}\right\} ,$$

where the first inequality follows from the convexity of $x \mapsto e^x$, while the second applies Hoeffding's lemma, as in (D.15) above. Then,

$$\mathbb{P}\left\{s \cdot \mathbf{e}_j^\top \left(\widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]\right) > \sqrt{\frac{2\log(27p_n^3)}{n-1}}\right\} \leqslant \frac{1}{27p^3} ,$$

and therefore taking a union bound over each $s \in \{\pm 1\}$ and each $j \in [p_n^2]$,

$$\mathbb{P}\left\{||\frac{1}{n}\sum_i \widehat{h}_1(X_i) - \mathbb{E}\left[h_1(X)\right]||_\infty > \sqrt{\frac{2\log(27p_n^3)}{n-1}}\right\} \leqslant \frac{1}{27p_n} .$$

Therefore, combining this with (D.16), with probability at least $1 - \frac{1}{9p_n}$,

$$||\widehat{\Sigma}_{h_1} - \Sigma_{h_1}||_\infty \leqslant 2\sqrt{\frac{80}{n-1} \cdot (1 + \log(27p_n^3))} + \sqrt{\frac{2\log(54p_n^3)}{n}} + 2\sqrt{\frac{2\log(27p_n^3)}{n-1}} \leqslant 100\sqrt{\frac{\log(p_n)}{n}} ,$$

where the last step uses the fact that $n, p_n \geqslant 2$. □

Proof of Lemma D.10. Since the statement is deterministic, we can treat $M \in \mathbb{R}^{p_n \times p_n}$ as fixed. Then

$$\begin{aligned}
&\mathsf{vec}(M)^\top \Sigma_{h_1} \mathsf{vec}(M) = \mathsf{Var}\left(\mathsf{vec}(M)^\top h_1(X)\right) \\
&= \mathsf{Var}\left(\mathsf{vec}(M)^\top \mathbb{E}[h(X, X') \mid X]\right) \\
&= \mathsf{Var}\left(\mathbb{E}[\mathsf{vec}(M)^\top h(X, X') \mid X]\right)
\end{aligned}$$

By the law of total variance,

$$\begin{aligned}
&\leqslant \mathsf{Var}\left(\mathsf{vec}(M)^\top h(X, X')\right) \\
&\leqslant \mathbb{E}\left[(\mathsf{vec}(M)^\top h(X, X'))^2\right] \\
&= \mathbb{E}\left[(\mathsf{vec}(M)^\top \left(\mathrm{sign}(X - X') \otimes \mathrm{sign}(X - X')\right))^2\right] \\
&= \mathbb{E}\left[\left(\mathrm{sign}(X - X')^\top M\, \mathrm{sign}(X - X')\right)^2\right]
\end{aligned}$$

Writing M_j as the jth column of M,

$$
\begin{aligned}
&= \mathbb{E}\left[\left(\sum_j \operatorname{sign}(X - X')^\top M_j \cdot \operatorname{sign}(X_j - X'_j)\right)^2\right] \\
&\leqslant \mathbb{E}\left[\left(\sum_j \left|\operatorname{sign}(X - X')^\top M_j\right|\right)^2\right] \\
&= \sum_{jk} \mathbb{E}\left[\left|\operatorname{sign}(X - X')^\top M_j\right| \cdot \left|\operatorname{sign}(X - X')^\top M_k\right|\right] \\
&\leqslant \sum_{jk} \sqrt{\mathbb{E}\left[\left|\operatorname{sign}(X - X')^\top M_j\right|^2\right]} \cdot \sqrt{\mathbb{E}\left[\left|\operatorname{sign}(X - X')^\top M_k\right|^2\right]} \\
&= \sum_{jk} \sqrt{M_j^\top \mathbb{E}\left[\operatorname{sign}(X - X')\operatorname{sign}(X - X')^\top\right] M_j} \cdot \sqrt{M_k^\top \mathbb{E}\left[\operatorname{sign}(X - X')\operatorname{sign}(X - X')^\top\right] M_k} \\
&= \sum_{jk} \sqrt{M_j^\top T M_j} \cdot \sqrt{M_k^\top T M_k} \\
&\leqslant \sum_{jk} \sqrt{||M_j||_2^2 \cdot \lambda_{\max}(T)} \cdot \sqrt{||M_k||_2^2 \cdot \lambda_{\max}(T)} \\
&= \lambda_{\max}(T) \cdot \left(\sum_j ||M_j||_2\right)^2 .
\end{aligned}
$$

Finally, by Wegkamp and Zhao (2013, Theorem 2.3), $\lambda_{\max}(T) \leqslant \lambda_{\max}(\Sigma)$. □

Lemma D.11. *Let $v \in \mathbb{R}^p$ be a fixed vector and let $Z_1, \ldots, Z_n \in [-1, 1]^p$ be random vectors, not necessarily independent, such that $v^\top(Z_i - \mathbb{E}[Z_i])$ is C-subgaussian for each i, that is,*

$$\mathbb{E}[\exp\{tv^\top(Z_i - \mathbb{E}[Z_i])\}] \leqslant \exp(Ct^2/2).$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_i \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2 \leqslant 20C(1 + \log(1/\delta)) .$$

Proof of Lemma D.11. For each i, by assumption,

$$\mathbb{E}\left[\exp\left\{t \cdot \frac{1}{\sqrt{C}} v^\top(Z_i - \mathbb{E}[Z_i])\right\}\right] \leqslant \exp\left\{\frac{t^2}{2}\right\} .$$

By Vershynin (2012, Lemma 5.5) (and tracking constants carefully in this Lemma), for each i,

$$\mathbb{E}\left[\exp\left\{\frac{1}{20C} \cdot \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2\right\}\right] \leqslant e .$$

By the convexity of $x \mapsto e^x$, then,

$$\mathbb{E}\left[\exp\left\{\frac{1}{20C} \cdot \frac{1}{n}\sum_i \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2\right\}\right] \leqslant \frac{1}{n}\sum_i \mathbb{E}\left[\exp\left\{\frac{1}{20C} \cdot \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2\right\}\right] \leqslant e .$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\left\{\frac{1}{n}\sum_i \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2 > t\right\} &\leqslant \mathbb{E}\left[\exp\left\{\frac{1}{20C}\frac{1}{n}\sum_i \left(v^\top(Z_i - \mathbb{E}[Z_i])\right)^2 - \frac{1}{20C}t\right\}\right] \\ &\leqslant \exp\left\{1 - \frac{1}{20C}t\right\} .\end{aligned}$$

Setting $t = 20C(1 + \log(1/\delta))$, then, we have proved the desired result.

□