Optimal Rates of Convergence of Transelliptical Component Analysis

Fang Han* and Han Liu†

May 10, 2013

Abstract

Han and Liu (2012) proposed a method named transelliptical component analysis (TCA) for conducting scale-invariant principal component analysis on high dimensional data with transelliptical distributions. The transelliptical family assumes that the data follow an elliptical distribution after unspecified marginal monotone transformations. In a double asymptotic framework where the dimension $d$ is allowed to increase with the sample size $n$, Han and Liu (2012) showed that one version of TCA attains a “nearly parametric” rate of convergence in parameter estimation when the parameter of interest is assumed to be sparse. This paper improves upon their results in two aspects: (i) Under the non-sparse setting (i.e., the parameter of interest is not assumed to be sparse), we show that a version of TCA attains the optimal rate of convergence up to a logarithmic factor; (ii) Under the sparse setting, we also lay out venues to analyze the performance of the TCA estimator proposed in Han and Liu (2012). In particular, we provide a “sign subgaussian condition” which is sufficient for TCA to attain an improved rate of convergence and verify a subfamily of the transelliptical distributions satisfying this condition.

Keyword: Transelliptical component analysis; Optimal rates of convergence; Double asymptotics; Minimax lower bound; Elliptical copula.

1 Introduction

Given $n$ observations $x_1, \ldots, x_n$ of a $d$ dimensional random vector $X \in \mathbb{R}^d$ with the covariance matrix $\Omega$, principal component analysis (PCA) estimates the top $m$ leading eigenvectors of $\Omega$. In the classical asymptotic theory where $d$ is fixed and $n$ goes to infinity, the

*Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA; e-mail: fhan@jhsph.edu

†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; e-mail: hanliu@princeton.edu
leading eigenvectors of the sample covariance matrix $S$ are efficient estimates (Anderson, 1958). However, under a double asymptotic framework where both $d$ and $n$ increase (with possibly $d \gg n$), Johnstone and Lu (2009) and Jung and Marron (2009) show that the leading eigenvector of $S$ could be an inconsistent estimate of the population quantity even when $X$ follows a Gaussian distribution. To handle this “curse of dimensionality”, one popular approach is to assume that the number of nonzero entries in the leading eigenvector of $\Omega$, denoted by $s$, is smaller than the sample size $n$. Under this assumption, different variants of sparse PCA have been developed (Jolliffe et al., 2003; d’Aspremont et al., 2004; Zou et al., 2006; Shen and Huang, 2008; Witten et al., 2009; Journée et al., 2010; Zhang and El Ghaoui, 2011) and the theoretical properties of these methods have been analyzed by Paul (2007); Ma (2013); Paul and Johnstone (2012); Vu and Lei (2012); Cai et al. (2012); Berthet and Rigollet (2013b); Shen et al. (2012); Berthet and Rigollet (2013a). For example, under the (sub)Gaussian model, Vu and Lei (2012) show that a version of sparse PCA achieves the $\sqrt{s \log d/n}$ rate of convergence in estimating the leading eigenvector of $\Omega$ and this rate is minimax optimal within a certain parameter space.

Very recently, Han and Liu (2012b) propose a new semiparametric generalization of sparse PCA. Their method, named transelliptical component analysis (TCA), builds upon the transelliptical distribution family. The transelliptical family assumes that, after unknown marginal transformations, the data follow an elliptical distribution. This family is closely related to the elliptical copula and contains many well known distributions, including multivariate Gaussian, rank-deficient Gaussian, multivariate-t, Cauchy, Kotz, logistic, etc.. TCA aims at conducting scale-invariant (sparse) PCA on the transelliptical data. Exploiting a transformed version of the Kendall’s tau correlation matrix, denoted by $\hat{\Sigma}$, Han and Liu (2012b) show that a version of TCA attains a “nearly parametric” rate of convergence $s \sqrt{\log d/n}$ in estimating the leading eigenvector of the latent generalized correlation matrix $\Sigma$ (detailed definitions will be provided in Section 2).

In this paper we extend the results in Han and Liu (2012b) in two directions. First, when the leading eigenvector of $\Sigma$ is not sparse, we advocate to directly use the leading eigenvector of $\hat{\Sigma}$ to estimate the corresponding population quantity. We provide explicit rate of convergence and show that this rate matches the minimax lower bound of PCA up to a logarithmic factor. Second, we establish a faster rate of convergence of the TCA estimator proposed by Han and Liu (2012b) using the optimization formulation adopted from Vu and Lei (2012). Though this algorithm is computationally intractable, its analysis provides deeper insights on the theoretical properties of TCA. More specifically, we provide a “sign subgaussian” condition which is sufficient for TCA to attain an improved rate of convergence $s \sqrt{\log d/n}$. This rate matches the minimax lower bound under the Gaussian model. Moreover, we verify that a subfamily of the transelliptical distributions satisfies this condition. Such a family contains all the continuous Gaussian copula distributions with block diagonal compound symmetry correlation matrix (detailed definitions will be
provided in Section 4). This is the first time such a result has been established.

In an independent work, Wegkamp and Zhao (2013) proposed to use the same transformed Kendall’s tau correlation coefficient estimator to analyze the elliptical copula factor model and proved a similar spectral norm convergence result as in Theorem 3.1 of this paper. The proof details are different and these two papers are independent work. In addition, the main focus of Wegkamp and Zhao (2013) is on analyzing the factor model while we focus on analyzing the properties of transelliptical component analysis.

The rest of this paper is organized as follows. In the next section we briefly overview transelliptical component analysis proposed by Han and Liu (2012b). In Section 3 we analyze the theoretical properties of TCA without the sparsity assumption. We show that the power method attains the minimax lower bound up to a logarithmic factor. In Section 4 we analyze the theoretical properties of TCA under the sparsity assumption. We show that under the “sign subgaussian” condition the TCA estimator proposed in Han and Liu (2012b) also attains an improved rate of convergence which matches the minimax lower bound under the Gaussian model. The technical proofs of these results are provided in Section 5. More discussions and conclusions are provided in Section 6.

2 Preliminaries and Background Overview

We briefly review the transelliptical distribution and transelliptical component analysis (TCA) proposed by Han and Liu (2012b). We start with an introduction of notation: Let \( M = [M_{ij}] \in \mathbb{R}^{d \times d} \) and \( \mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d \). We denote \( \mathbf{v}_I \) to be the subvector of \( \mathbf{v} \) whose entries are indexed by a set \( I \). We also denote \( M_{I, J} \) to be the submatrix of \( M \) whose rows are indexed by \( I \) and columns are indexed by \( J \). Let \( M_I \) and \( M_J \) be the submatrix of \( M \) with rows indexed by \( I \), and the submatrix of \( M \) with columns indexed by \( J \). Let \( \text{supp}(\mathbf{v}) := \{ j : v_j \neq 0 \} \). For \( 0 < q < \infty \), we define the \( \ell_0 \), \( \ell_q \) and \( \ell_{\infty} \) vector norms as

\[
\| \mathbf{v} \|_0 := \text{card}(\text{supp}(\mathbf{v})), \quad \| \mathbf{v} \|_q := \left( \sum_{i=1}^{d} |v_i|^q \right)^{1/q} \quad \text{and} \quad \| \mathbf{v} \|_{\infty} := \max_{1 \leq i \leq d} |v_i|.
\]

Let \( \lambda_j(M) \) be the \( j \)-th largest eigenvalue of \( M \) and \( \Theta_j(M) \) be the corresponding leading eigenvector. In particular, we let \( \lambda_{\max}(M) := \lambda_1(M) \). We define \( \mathbb{S}^{d-1} := \{ \mathbf{v} \in \mathbb{R}^d : \| \mathbf{v} \|_2 = 1 \} \) to be the \( d \)-dimensional unit sphere. We define the matrix \( \ell_{\max} \) norm and \( \ell_2 \) norm as \( \| M \|_{\max} := \max \{ |M_{ij}| \} \) and \( \| M \|_2 := \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \| M \mathbf{v} \|_2 \). We define \( \text{diag}(M) \) to be a diagonal matrix with \( [\text{diag}(M)]_{ij} = M_{ij} \) for \( j = 1, \ldots, d \). We also denote \( \text{vec}(M) := (M_{11}^T, \ldots, M_{dd}^T)^T \). For any two vector \( \mathbf{a}, \mathbf{b} \in \mathbb{R}^d \), we denote \( \langle \mathbf{a}, \mathbf{b} \rangle := \mathbf{a}^T \mathbf{b} \) and \( \text{sign}(\mathbf{a}) := (\text{sign}(a_1), \ldots, \text{sign}(a_d))^T \), where \( \text{sign}(x) = x/|x| \) with the convention \( 0/0 = 0 \).
2.1 Transelliptical Component Analysis

Transelliptical component analysis builds upon the concepts of transelliptical distribution and elliptical distribution. We provide a definition of the elliptical distribution using the stochastic representation as in Fang et al. (1990). In the sequel, for any two random vectors $X$ and $Y$, we denote $X \overset{d}{=} Y$ if they are identically distributed.

**Definition 2.1** (Fang et al. (1990)). A random vector $Z = (Z_1, \ldots, Z_d)^T$ follows an elliptical distribution if and only if $Z$ has a stochastic representation: $Z \overset{d}{=} \mu + \xi A U$. Here $\mu \in \mathbb{R}^d$, $q := \text{rank}(A) \in \mathbb{R}^{d \times q}$, $\xi \geq 0$ is a random variable independent of $U$, $U \in S_{q-1}$ is uniformly distributed on the unit sphere in $\mathbb{R}^q$. In this setting, letting $\Sigma := AA^T$, we denote $Z \sim EC_d(\mu, \Sigma, \xi)$. Here $\Sigma$ is called the scatter matrix.

The elliptical family can be viewed as a semiparametric generalization of the Gaussian family, maintaining the symmetric property of the Gaussian distribution but allowing heavy-tails and tail-dependence. The transelliptical distribution family further relaxes the symmetric assumption of the elliptical distribution by assuming that, after unspecified strictly increasing marginal transformations, the data are elliptically distributed. A formal definition of the transelliptical distribution is as follows.

**Definition 2.2** (Han and Liu (2012b)). A continuous random vector $X = (X_1, \ldots, X_d)^T$ follows a transelliptical distribution, denoted by $X \sim TE_d(\Sigma, \xi; f_1, \ldots, f_d)$, if there exist univariate strictly increasing functions $f_1, \ldots, f_d$ such that

$$(f_1(X_1), \ldots, f_d(X_d))^T \sim EC_d(0, \Sigma, \xi), \text{ where } \text{diag}(\Sigma) = I_d.$$  

Here $I_d \in \mathbb{R}^{d \times d}$ is the $d$-dimensional identity matrix and $\Sigma$ is called the latent generalized correlation matrix.

We note that the transelliptical distribution is closely related to the nonparanormal distribution (Liu et al., 2009, 2012a; Xue and Zou, 2012; Han and Liu, 2012a; Han et al., 2013) and meta-elliptical distribution (Fang et al., 2002). The nonparanormal distribution assumes that after unspecified strictly increasing marginal transformations, the data are Gaussian distributed. It is easy to see that the transelliptical family contains the nonparanormal family. On the other hand, it is subtle to elaborate the difference between the transelliptical and meta-elliptical. In short, the transelliptical family contains meta-elliptical family. Compared to the meta-elliptical, the transelliptical family does not require the random vectors to have densities and brings new insight into both theoretical analysis and model interpretability. We refer to Liu et al. (2012b) for more detailed discussion on the comparison between the transelliptical family, nonparanormal, and meta-elliptical families.

TCA exploits the transelliptical distribution family and can be viewed as a scale-invariant variant of (sparse) PCA on the latent elliptically distributed random vector...
\( f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \). In particular, TCA aims at recovering the leading eigenvector \( \theta_1 := \Theta_1(\Sigma) \) of the latent generalized correlation matrix \( \Sigma \). When \( d < n \), we do not need extra assumptions on \( \theta_1 \). In high dimensions when \( d > n \), we assume that the leading eigenvector of \( \Sigma \) is sparse. In this setting, the statistical model exploited in Han and Liu (2012b) is as follows.

**Definition 2.3** (Han and Liu (2012b)). The TCA model, denoted by \( \mathcal{M}_d(\Sigma, \xi, f, s) \), is defined to be the set of distributions:

\[
\mathcal{M}_d(\Sigma, \xi, f, s) := \{ X : X \sim T\mathcal{E}_d(\Sigma, \xi; f_1, \ldots, f_d) \text{ such that } \| \Theta_1(\Sigma) \|_0 \leq s. \}
\]

(2.1)

where \( f := \{ f_j \}_{j=1}^d \) is a set of univariate strictly increasing functions.

By treating both the generating variable \( \xi \) and the marginal transformation functions \( f = \{ f_j \}_{j=1}^d \) as nuisance parameters, Han and Liu (2012b) propose to use a transformed Kendall’s tau correlation matrix to estimate the latent generalized correlation matrix \( \Sigma \). More specifically, letting \( x_1, \ldots, x_n \) be \( n \) independent and identically distributed observations of a random vector \( X \in T\mathcal{E}_d(\Sigma, \xi; f_1, \ldots, f_d) \), the Kendall’s tau correlation coefficient between the variables \( X_j \) and \( X_k \) is defined as

\[
\hat{\tau}_{jk} := \frac{2}{n(n-1)} \sum_{i<i'} \text{sign}((x_{ij} - x_{ij'}) (x_{ik} - x_{ik'})).
\]

Its population quantity can be written as

\[
\tau_{jk} := \mathbb{P}((X_j - \bar{X}_j)(X_k - \bar{X}_k) > 0) - \mathbb{P}((X_j - \bar{X}_j)(X_k - \bar{X}_k) < 0),
\]

(2.2)

where \( \bar{X} = (\bar{X}_1, \ldots, \bar{X}_d)^T \) is an independent copy of \( X \). For the transelliptical family, it is well known that \( \Sigma_{jk} = \sin(\frac{\pi}{2} \tau_{jk}) \) (Also see Theorem 3.1 in Han and Liu (2012b)). A latent generalized correlation matrix estimator \( \hat{\Sigma} := [\hat{\Sigma}_{jk}] \), called the transformed Kendall’s tau matrix, is then defined as:

\[
\hat{\Sigma}_{jk} = \sin \left( \frac{\pi}{2} \hat{\tau}_{jk} \right).
\]

(2.3)

Once \( \hat{\Sigma} \) is calculated, the TCA estimator \( \hat{\theta}_1 \) of \( \theta_1 \), is defined as

\[
\hat{\theta}_1 := \arg \max_{v \in \mathbb{R}^d} v^T \hat{\Sigma} v, \quad \text{subject to } \| v \|_2 = 1 \text{ and } \| v \|_0 \leq s.
\]

(2.4)

The optimization formulation shown in Equation (2.4) is initiated in Vu and Lei (2012). Although it is combinatoric in nature and hard to compute, this formulation helps understanding the theoretical properties of TCA. Han and Liu (2012b) show that, when \( X \in \mathcal{M}_d(\Sigma, \xi, f, s) \), the estimator \( \hat{\theta}_1 \) satisfies that

\[
| \sin \angle(\hat{\theta}_1, \theta_1) | = O_P \left( \frac{1}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \cdot s \sqrt{\frac{\log d}{n}} \right),
\]

(2.5)
where \( \theta_1 := \Theta_1(\Sigma) \) and \( \angle(\hat{\theta}_1, \theta_1) \) is the angle between \( \hat{\theta}_1 \) and \( \theta_1 \). Accordingly, compared to the results in Vu and Lei (2012), Han and Liu (2012b) call the rate in (2.5) “nearly parametric”.

3 TCA Theory: Non-Sparse Setting

In this section we provide the rate of convergence of \( \Theta_1(\hat{\Sigma}) \) to \( \theta_1 \) without the sparsity assumption. The next theorem shows that, under the transelliptical distribution family, the convergence rate of \( \hat{\Sigma} \) to \( \Sigma \) under the spectral norm is \( \sqrt{d \log d/n} \), which is the same parametric rate as in Vershynin (2010); Lounici (2012); Bunea and Xiao (2012) when there is not any additional structure. For notational simplicity, in the sequel we assume that the sample size \( n \) is even. When \( n \) is odd, we can always use \( n - 1 \) data points without affecting the obtained rate of convergence.

**Theorem 3.1.** Let \( x_1, \ldots, x_n \) be \( n \) observations of \( X \sim TE_d(\Sigma, \xi; f_1, \ldots, f_d) \). Let \( \hat{\Sigma} \) be the transformed Kendall’s tau correlation matrix defined in Equation (2.3). We have, for \( n \) large enough and \( 0 < \alpha < 1 \), with probability larger than \( 1 - 2\alpha - \alpha^2 \),

\[
\|\hat{\Sigma} - \Sigma\|_2 \leq 2\pi^2 \sqrt{\frac{\{d\|T\|_2 + \|T\|_2^2\} \log(d/\alpha)}{3n} + \pi^2 \cdot \frac{d \log(d/\alpha)}{n}},
\]

where \( T := [T_{jk}] \in \mathbb{R}^{d \times d} \) with \( T_{jk} := \tau_{jk} \).

Theorem 3.1 indicates that, when \( \|T\|_2 < C \) for some constant \( C \) and \( d \log d/n \to 0 \), \( \|\hat{\Sigma} - \Sigma\|_2 = O_p(\sqrt{d \log d/n}) \). By the discussion of Theorem 2 in Lounici (2012), when \( \|T\|_2 \) is upper bounded by a fixed constant, the rate of convergence is minimax optimal up to a logarithmic factor with respect to a suitable parameter space. Here we note that Equation (3.1) can be rewritten as

\[
\|\hat{\Sigma} - \Sigma\|_2 \leq 2\pi^2 \|T\|_2 \sqrt{\frac{\{d\|T\|_2 + 1\} \log(d/\alpha)}{3n} + \pi^2 \cdot \frac{d \log(d/\alpha)}{n}},
\]

where the term \( d/\|T\|_2 = \text{Tr}(T)/\|T\|_2 \) is called the effective rank of \( T \). For more details, see Vershynin (2010); Lounici (2012); Bunea and Xiao (2012).

Built on Theorem 3.1, the next corollary provides an explicit rate of convergence for

**Corollary 3.2.** Under the conditions of Theorem 3.1, we have, with probability larger than \( 1 - 2\alpha - \alpha^2 \),

\[
|\sin \angle(\Theta_1(\hat{\Sigma}), \theta_1)| \leq \frac{2\pi^2}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \left[ 2\sqrt{\frac{\{d\|T\|_2 + \|T\|_2^2\} \log(d/\alpha)}{3n} + \frac{d \log(d/\alpha)}{n}} \right],
\]
Corollary 3.2 can be proved by combining Theorem 3.1 and a part of the proof in Theorem 2.2 in Vu and Lei (2012) (an explicit version can be found in Lemma 4.3 in Wang et al. (2013)). We accordingly omit the proof here. Built upon Theorems 3.1 and Corollary 3.2, we see that $\Theta_1(\hat{\Sigma})$ is a consistent estimator of $\theta_1$ under multiple settings, for example when $\lambda_1(\Sigma) - \lambda_2(\Sigma)$ and $\|T\|_2$ do not scale with $(n, d, s)$ and $d \log d/n \to 0$. However, when $d > n$, $\Theta_1(\hat{\Sigma})$ may not be a consistent estimator. We will accordingly discuss the sparse PCA version of TCA in the next section.

4 TCA Theory: Sparse Setting

In this section we analyze the theoretical properties of TCA with the sparsity assumption. More specifically, we provide sufficient conditions under which the TCA estimator $\hat{\theta}_1$ defined in Equation (2.4) attains the optimal $\sqrt{s \log d/n}$ rate of convergence.

From Equation (2.5), we have that $|\sin \angle(\hat{\theta}_1, \theta_1)| = O_P(s \sqrt{\log d/n})$. Our main focus in the rest of this section is to characterize sufficient conditions under which we can get an improved rate of convergence:

$$|\sin \angle(\hat{\theta}_1, \theta_1)| = O_P(\sqrt{s \log d/n}).$$

Such a rate has been shown to be minimax optimal under the Gaussian model. Obtaining such an improved rate is technically challenging since the data could be very heavy-tailed and the transformed Kendall’s correlation matrix has a much more complex structure than the Person’s covariance/correlation matrix.

In the following we lay out a venue to analyze the statistical efficiency of $\hat{\theta}_1$. In particular, we characterize a subset of the transelliptical distributions for which $\hat{\theta}_1$ can approximate $\theta_1$ in an improved rate as in (4.1). More specifically, we provide a “sign subgaussian” condition which is sufficient for TCA to attain the rate in (4.1). This condition is related to the subgaussian assumption in Vu and Lei (2012); Lounici (2012); Bunea and Xiao (2012) (see Assumption 2.2 in Vu and Lei (2012), for example). Before proceeding to the formal definition of this condition, we first define the operator $\psi : \mathbb{R} \to \mathbb{R}$ as follows:

**Definition 4.1.** For any random variable $Y \in \mathbb{R}$, the operator $\psi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\psi(Y; \alpha, t_0) := \inf \{ c > 0 : \mathbb{E} \exp \{ t(Y^\alpha - \mathbb{E}Y^\alpha) \} \leq \exp(ct^2), \text{ for } |t| < t_0 \}.$$  (4.2)

The operator $\psi(\cdot)$ can be used to quantify the tail behaviors of random variables. We recall that a zero-mean random variable $X \in \mathbb{R}$ is said to be subgaussian distributed if there exists a constant $c$ such that $\mathbb{E} \exp(tX) \leq \exp(ct^2)$ for all $t \in \mathbb{R}$. A zero-mean random variable $Y \in \mathbb{R}$ with $\psi(Y; 1, \infty)$ bounded is well known to be a subgaussian distribution, which implies a tail probability

$$\mathbb{P}(|Y - \mathbb{E}Y| > t) < 2 \exp(-t^2/(4c)).$$
where $c$ is the constant defined in Equation (4.2). Moreover, $\psi(Y; \alpha, t_0)$ is related to the Orlicz $\psi_2$-norm. A formal definition of the Orlicz norm is provided as follows.

**Definition 4.2.** For any random variable $Y \in \mathbb{R}$, its Orlicz $\psi_2$-norm is defined as

$$
\|Y\|_{\psi_2} := \inf \left\{ c > 0 : \mathbb{E} \exp \left( \frac{|Y|}{c}^2 \right) \leq 2 \right\}.
$$

It is well known that a random variable $Y$ has $\psi(Y; 1, \infty)$ to be bounded if and only if $\|Y\|_{\psi_2}$ in Definition 4.2 is bounded (van de Geer and Lederer, 2011). We refer to Lemma A.1 in the appendix for a more detailed description on this property.

Another relevant norm to $\psi(\cdot)$ is the subgaussian norm $\|\cdot\|_{\phi_2}$ defined in Vershynin (2010). A former definition of the subgaussian norm is as follows.

**Definition 4.3.** For any random variable $X \in \mathbb{R}$, its subgaussian norm is defined as

$$
\|X\|_{\phi_2} := \sup_{k \geq 1} k^{-1/2} \left( \mathbb{E}|X|^k \right)^{1/k}.
$$

The subgaussian norm is also highly related to the subgaussian distribution. In particular, we have if $\mathbb{E}X = 0$, then $\mathbb{E}\exp(tX) \leq \exp(Ct^2\|X\|_{\phi_2}^2)$. We say that a random vector $X \in \mathbb{R}^d$ is subgaussian distributed if for any vector $v \in \mathbb{R}^d$, $v^T X$ is subgaussian distributed. We refer to Vershynin (2010) for more details on the properties of the subgaussian norm.

Using the operator $\psi(\cdot)$, we now proceed to define the sign subgaussian condition. Here for any vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, we remind that $\text{sign}(v) := (\text{sign}(v_1), \ldots, \text{sign}(v_d))^T$.

**Definition 4.4 (Sign subgaussian condition).** For a random vector $X = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$, let $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)^T$ be an independent copy of $X$, $X$ satisfies the sign subgaussian condition if and only if

$$
\sup_{v \in S^{d-1}} \psi \left( \left\langle \text{sign}(X - \widetilde{X}), v \right\rangle ; 2, t_0 \right) \leq K(\|\Sigma\|_2 + \|T\|_2)^2,
$$

where $K$ is a fixed constant and $t_0$ is a positive number such that $t_0K(\|\Sigma\|_2 + \|T\|_2)^2$ is lower bounded by a fixed constant.

To gain more insights about the sign subgaussian condition, In the following we point out two distribution families of interest that satisfy the sign subgaussian condition.

**Proposition 4.5.** For any random vector $X \in \mathbb{R}^d$ and $\widetilde{X}$ being an independent copy of $X$, if we have

$$
\sup_{v \in S^{d-1}} \left\| \left\langle \text{sign}(X - \widetilde{X}), v \right\rangle^2 - v^T T v \right\|_{\psi_2} \leq L_1(\|\Sigma\|_2 + \|T\|_2),
$$

where $T := [T_{jk}] \in \mathbb{R}^{d \times d}$ with $T_{jk} := \tau_{jk}$ and $L_1$ is a fixed constant, then $X$ satisfies the sign subgaussian condition by setting $t_0 = \infty$ and $K = 5L_1^2/2$ in Equation (4.3).
Proposition 4.6. For any random vector \( X \in \mathbb{R}^d \) and \( \tilde{X} \) being an independent copy of \( X \), if there exists an absolute constant \( L_2 \) such that \( \operatorname{sign}(X - \tilde{X}) \) is subgaussian distributed and

\[
\|v^T \operatorname{sign}(X - \tilde{X})\|_2^2 \leq \frac{L_2 \|\Sigma\|_2^2}{2} \quad \text{for all } v \in \mathbb{S}^{d-1},
\]  

then \( X \) satisfies the sign subgaussian condition with \( t_0 = c(\|\Sigma\|_2 + \|T\|_2)^{-1} \) and \( K = C \) in Equation (4.3), where \( c \) and \( C \) are two fixed absolute constants.

Proposition 4.6 builds a bridge between the sign subgaussian condition and the Assumption 1 in Bunea and Xiao (2012) and Lounici (2012). More specifically, \( X \) satisfies Equation (4.5) is equivalent to saying that \( \operatorname{sign}(X - \tilde{X}) \) satisfies the sign subgaussian condition defined in Bunea and Xiao (2012). Therefore, Proposition 4.6 can be treated as an explanation of why we call the condition in Equation (4.3) “sign subgaussian”. However, by Lemma 5.14 in Vershynin (2010), the sign subgaussian condition is weaker than that of Equation (4.5), i.e., a distributions satisfying the sign subgaussian condition does not necessarily satisfy the condition in Equation (4.5).

The sign subgaussian condition is intuitive due to its relationship with the Orlicz and subgaussian norms. However, it is extremely difficult to verify whether a given distribution satisfies this condition. The main difficulty lies in the fact that we must sharply characterize the tail behavior of the summation of a sequence of possibly correlated discrete random variables, which is much harder than analyzing the summation of Gaussian random variables as usually done in the literature.

In the following we provide several examples that satisfy the sign subgaussian condition. The next theorem shows that for any transelliptically distributed random vector \( X \sim \mathcal{T}E_d(\Sigma, \xi; f_1, \ldots, f_d) \) such that \( f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(0, I_d) \), the distribution of \( X \) satisfies the condition shown in Equation (4.3). The proof of Theorem 4.7 is in Section 5.3, from which see that even for such a simple distribution, the analysis is nontrivial.

**Theorem 4.7.** Suppose that \( X \sim \mathcal{T}E_d(\Sigma, \xi; f_1, \ldots, f_d) \) is transelliptically distributed such that \( f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(0, I_d) \). Then \( X \) satisfies the sign subgaussian condition shown in Equation (4.3).

In the next theorem, we provide a stronger version of Theorem 4.7. We call a square matrix compound symmetric if the off diagonal values of the matrix is equal. The next theorem shows that for any transelliptically distributed random vector \( X \) such that \( f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(0, \Sigma) \) with \( \Sigma \) a restricted compound symmetric matrix satisfies Equation (4.5), and therefore satisfies the sign subgaussian condition.

**Theorem 4.8.** Suppose that \( X \sim \mathcal{T}E_d(\Sigma, \xi; f_1, \ldots, f_d) \) is transelliptically distributed such that \( f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(0, \Sigma) \), with \( \Sigma \) a compound symmetric matrix (i.e., \( \Sigma_{jk} = \rho \) for all \( j \neq k \)). Then if \( 0 \leq \rho := \Sigma_{12} \leq C_0 < 1 \) for some absolute
positive constant $C_0$, we have that $X$ satisfies Equation (4.5), and therefore satisfies the sign subgaussian condition shown in Equation (4.3).

Although Theorem 4.7 can be directly proved using the result in Theorem 4.8, the proof of Theorem 4.7 contains utterly different techniques which are more transparent and illustrate the main challenges of analyzing binary sequences even in the uncorrelated setting. Therefore, we still list this theorem separately and provide a separate proof in Section 5.3. Theorem 4.8 leads to the following corollary, which characterizes a subfamily of the transelliptical distributions satisfying the sign subgaussian condition.

**Corollary 4.9.** Suppose that $X \sim TE_d(\Sigma, \xi; f_1, \ldots, f_d)$ is transelliptically distributed such that $f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N_d(0, \Sigma)$ with $\Sigma$ a block diagonal compound symmetric matrix, i.e.,

$$
\Sigma = \begin{pmatrix}
\Sigma_1 & 0 & 0 & \cdots & 0 \\
0 & \Sigma_2 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \cdots & \vdots \\
0 & 0 & 0 & \cdots & \Sigma_q
\end{pmatrix},
$$

(4.6)

where $\Sigma_k \in \mathbb{R}^{d_k \times d_k}$ for $k = 1, \ldots, q$ is compound symmetric matrix with $\rho_k := \|\Sigma_k\|_2 \geq 0$. We have, if $q$ is upper bounded by an absolute positive constant and $0 \leq \rho_k \leq C_1 < 1$ for some absolute positive constant $C_1$, then $X$ satisfies the sign subgaussian condition.

We call the matrix in the form of Equation (4.6) block diagonal compound symmetric matrix. Corollary 4.9 indicates that any transelliptical distribution with a latent Gaussian distribution and block diagonal compound symmetric latent generalized correlation matrix satisfies the sign subgaussian condition.

Using the sign subgaussian condition, we have the following main result, which shows that for the transelliptical distributed random vector $X \in M_d(\Sigma, \xi, f, s)$ satisfying the condition in Definition 4.4, the TCA estimator $\hat{\theta}_1$ can approximate $\theta_1$ in an improved rate of convergence shown in (4.1).

**Theorem 4.10.** Let $\hat{\theta}_1$ be the global optimum to Equation (2.4) and $x_1, \ldots, x_n$ be $n$ observations of $X \sim M_d(\Sigma, \xi, f, s)$, such that $X := (f_1(X_1), \ldots, f_d(X_d))^T$ satisfies the sign subgaussian condition. We have, for $n$ large enough, with probability larger than $1 - 2\alpha - \alpha^2$,

$$
|\sin \angle(\hat{\theta}_1, \theta_1)| \leq \frac{4\pi^2}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \left( (\|\Sigma\|_2 + \|T\|_2)(2K)^{1/2} \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}} + \frac{s \log(d/\alpha)}{n} \right),
$$

where $K$ is the fixed constant defined in Equation (4.3).

The rate of convergence presented in Theorem 4.10 shows that under various settings the rate of convergence for $\hat{\theta}_1$ is $O_P(\sqrt{s \log d/n})$, which is the parametric and minimax optimal rate shown in Vu and Lei (2012) within the Gaussian family.
5 Technical Proofs

In this section we provide the technical proofs of the theorems shown in Sections 3 and 4.

5.1 Proof of Theorem 3.1

To prove Theorem 3.1, we first present a useful lemma. We remind that the population and empirical Kendall’s tau correlation matrices are defined as \( T \) and \( \hat{T} \) with \( T_{jk} = \tau_{jk} \) and \( \hat{T}_{jk} = \hat{\tau}_{jk} \). The next lemma provides an upper bound on the spectral norm of \( \hat{T} - T \).

**Lemma 5.1.** Under the conditions in Theorem 3.1, whenever

\[
n \geq \frac{64(d\|T\|_2 + \|\hat{T}\|_2)\log(d/\alpha)}{3},
\]

we have, with probability larger than \( 1 - 2\alpha \),

\[
\|\hat{T} - T\|_2 \leq 4\sqrt{\frac{d\|T\|_2 + \|\hat{T}\|_2^2}{3n}} \log\left(\frac{d}{\alpha}\right).
\]

**(5.1)**

**Proof.** Reminding that \( x_i := (x_{i1}, \ldots, x_{id})^T \), for \( i \neq i' \), let

\[
S_{i,i'} := (\text{sign}(x_{i1} - x_{i'1}), \ldots, \text{sign}(x_{id} - x_{i'd}))^T.
\]

We denote \( \hat{\Delta}_{i,i'} \) to be \( n(n-1) \) random matrices with

\[
\hat{\Delta}_{i,i'} := \frac{1}{n(n-1)}(S_{i,i'}S_{i,i'}^T - T).
\]

By simple calculation, we have \( \hat{T} - T = \sum_{i,i'} \hat{\Delta}_{i,i'} \) and \( \hat{T} - T \) is a U-statistic.

In the following we extends the standard decoupling trick from Hoeffding (1963) from the U-statistic of random variables to the matrix setting. The extension relies on the matrix version of the Laplace transform method. For any square matrix \( M \in \mathbb{R}^d \), we define

\[
\exp(M) := I_d + \sum_{k=1}^{\infty} \frac{M^k}{k!},
\]

where \( k! \) represents the factorial product of \( k \). Using Proposition 3.1 in Tropp (2010), we have

\[
\mathbb{P}[\lambda_{\text{max}}(\hat{T} - T) \geq \ell] \leq \inf_{\theta > 0} e^{-\theta \ell} \mathbb{E}\left[\text{Tr} e^{\theta(\hat{T} - T)}\right],
\]

**(5.2)**

and we bound \( \mathbb{E}\left[\text{Tr} e^{\theta(\hat{T} - T)}\right] \) as follows.

The trace exponential function

\[
\text{Tr exp} : A \rightarrow \text{Tr} A
\]
is a convex mapping from the space of self-adjoint matrix to \( \mathbb{R}^+ \) (see Section 2.4 of Tropp (2010) and reference therein). Let \( m = n/2 \). For any permutation \( \sigma \) of \( 1, \ldots, n \), let \( (i_1, \ldots, i_n) := \sigma(1, \ldots, n) \). For \( r = 1, \ldots, m \), we define \( S_\sigma^r \) and \( \hat{\Delta}_\sigma^r \) to be

\[
S_\sigma^r := S_{2i, 2i-1} \quad \text{and} \quad \hat{\Delta}_\sigma^r := \frac{1}{m} (S_\sigma^r [S_\sigma^r]^T - T).
\]

Moreover, for \( i = 1, \ldots, m \), let

\[
S_i := S_{2i, 2i-1} \quad \text{and} \quad \hat{\Delta}_i := \frac{1}{m} (S_i S_i^T - T).
\]

The convexity of the trace exponential function implies that

\[
\text{Tr} e^{\theta (\hat{T} - T)} = \text{Tr} e^{\theta \sum_{i,i'} \hat{\Delta}_{i,i'}}
\]

\[
= \text{Tr} \exp \left\{ \frac{1}{\text{card}(S_n)} \sum_{\sigma \in S_n} \theta \sum_{r=1}^m \hat{\Delta}_\sigma^r \right\}
\]

\[
\leq \frac{1}{\text{card}(S_n)} \sum_{\sigma \in S_n} \text{Tr} e^{\theta \sum_{r=1}^m \hat{\Delta}_\sigma^r},
\]

(5.3)

where \( S_n \) is the permutation group of \( \{1, \ldots, n\} \). Taking expectation on both sides of Equation (5.3) gives that

\[
E \text{Tr} e^{\theta (\hat{T} - T)} \leq E \text{Tr} e^{\theta \sum_{i=1}^m \hat{\Delta}_i}.
\]

(5.4)

According to the definition, \( \hat{\Delta}_1, \ldots, \hat{\Delta}_m \) are \( m \) independent and identically distributed random matrices, and this finishes the decoupling step.

Combing Equations (5.2) and (5.4), we have

\[
P[\lambda_{\text{max}}(\hat{T} - T) \geq t] \leq \inf_{\theta > 0} e^{-\theta t} E \text{Tr} e^{\theta \sum_{i=1}^m \hat{\Delta}_i}.
\]

(5.5)

Recall that \( E \hat{\Delta}_i = 0 \). Following the proof of Theorem 6.1 in Tropp (2010), if we can show that there are some positive numbers \( R_1 \) and \( R_2 \) such that

\[
\lambda_{\text{max}}(\hat{\Delta}_i) \leq R_1, \quad \| \sum_{i=1}^m E \hat{\Delta}_i^2 \|_2 \leq R_2,
\]

then the right hand side of Equation (5.5) can be bounded by

\[
\inf_{\theta > 0} e^{-\theta t} E \text{Tr} e^{\theta \sum_{i=1}^m \hat{\Delta}_i} \leq d \exp \left\{ -\frac{t^2/2}{R_2 + R_1 t/3} \right\}.
\]

We first show that \( R_1 = \frac{2d}{m} \). Because \( \| \hat{\Delta}_i \|_{\text{max}} \leq 2/m \), by simple calculation, we have

\[
\lambda_{\text{max}}(\hat{\Delta}_i) \leq \| \hat{\Delta}_i \|_1 \leq d \cdot \| \hat{\Delta}_i \|_{\text{max}} \leq \frac{2d}{m}.
\]
We then calculate $R^2$. For this, we have

$$\sum_{i=1}^{m} E \hat{\Delta}_i^2 = \frac{1}{m} E(S_1 S_1^T - T)^2 = \frac{1}{m} (E(dS_1 S_1^T) - T^2) = \frac{1}{m} (dT - T^2).$$

Accordingly,

$$\| \sum_{i=1}^{m} E \hat{\Delta}_i^2 \|_2 \leq \frac{1}{m} (d\|T\|_2 + \|T\|_2^2),$$

so we have $R^2 = \frac{1}{m} (d\|T\|_2 + \|T\|_2^2)$.

Thus using Theorem 6.1 in Tropp (2010), for any $t \leq R^2/R_1 = \frac{d\|T\|_2 + \|T\|_2^2}{2d}$, we have

$$P\left\{ \lambda_{\max}(\hat{T} - T) \geq t \right\} \leq d \cdot \exp\left( -\frac{3nt^2}{16(d\|T\|_2 + \|T\|_2^2)} \right).$$

A similar argument holds for $\lambda_{\max}(-\hat{T} + T)$. Accordingly, we have

$$P\left\{ \|\hat{T} - T\|_2 \geq t \right\} \leq 2d \cdot \exp\left( -\frac{3nt^2}{16(d\|T\|_2 + \|T\|_2^2)} \right).$$

Moreover, because $\|T\|_2 \geq 1$, $t \leq 1/2$ implies $t \leq \frac{d\|T\|_2 + \|T\|_2^2}{2d}$. This completes the proof.  

To prove the main theorem, we still need one more lemma, which connects $\sqrt{1 - \Sigma_{jk}^2}$ to a Gaussian distributed random vector $(X,Y)^T \in \mathbb{R}^2$ and plays a key role in bounding $\|\hat{\Sigma} - \Sigma\|_2$ by $\|\hat{T} - T\|_2$.

**Lemma 5.2.** Provided that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(0, \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix} \right),$$

and $\sigma := \sin(\frac{\pi}{2} \tau)$ with $\tau$ the Kendall’s tau correlation coefficient of $X,Y$, we have

$$E|XY| = EXY \text{sign}(XY) + \frac{2}{\pi} \sqrt{1 - \sigma^2}.$$

**Proof.** Without loss of generality, assume that $\sigma > 0, \tau > 0$ (otherwise show for $-Y$ instead of $Y$). Define

$$\beta_+ = E|XY| I(XY > 0), \quad \beta_- = E|XY| I(XY < 0),$$

where $I(\cdot)$ is the indicator function. We then have

$$E|XY| = \beta_+ + \beta_-, \quad EXY = \sigma = \beta_+ - \beta_-.$$

(5.6)
To compute $\beta_+$, using the fact that

$$X \overset{d}{=} \sqrt{\frac{1+\sigma}{2}} Z_1 + \sqrt{\frac{1-\sigma}{2}} Z_2, \quad Y \overset{d}{=} \sqrt{\frac{1+\sigma}{2}} Z_1 - \sqrt{\frac{1-\sigma}{2}} Z_2,$$

where $Z_1, Z_2 \sim N_1(0,1)$ are independently and identically distributed.

Let $F_{X,Y}$ and $F_{Z_1,Z_2}$ be joint distribution functions of $(X,Y)^T$ and $(Z_1,Z_2)^T$. We have

$$\beta_+ = \int_{xy>0} |xy| dF_{X,Y}(x,y) = \int_{xy>0} \frac{(x+y)^2 - (x-y)^2}{4} dF_{X,Y}(x,y) = \int_{z_1^2>\frac{1+\sigma}{1-\sigma} \frac{1-\sigma}{2} z_2^2} dF_{Z_1,Z_2}(z_1,z_2) = \int_{0}^{+\infty} \int_{-\alpha}^{\alpha} 2 \left\{ \frac{1+\sigma}{2} r^2 \cos^2(\theta) - \frac{1-\sigma}{2} r^2 \sin^2(\theta) \right\} \cdot \frac{1}{2\pi} e^{-r^2/2} r d\theta dr,$$

where $\alpha := \arcsin \left(\sqrt{\frac{1+\sigma}{2}}\right)$. By simple calculation, we have

$$\int_{0}^{+\infty} r^3 e^{-r^2/2} dr = \frac{1}{2} \int_{0}^{\infty} u e^{-u/2} du = 2.$$

Accordingly, we can proceed the proof and show that

$$\beta_+ = \int_{0}^{+\infty} \int_{-\alpha}^{\alpha} (\cos(2\theta) + \sigma) \cdot r^3 \frac{1}{2\pi} e^{-r^2/2} d\theta dr = \frac{1}{\pi} \left( \sin(2\alpha) + 2\alpha \sigma \right).$$

(5.7)

Since $\sin(2\alpha) = \sqrt{1-\sigma^2} = \cos(\pi/2)$ and $\alpha \geq \arcsin(\sqrt{1/2}) \geq \pi/4$, we have that $2\alpha = \frac{\pi}{2}(1+\tau)$, and then Equation (5.7) continues to give

$$\beta_+ = \frac{\sigma}{2} (1+\tau) + \frac{1}{\pi} \sqrt{1-\sigma^2}.$$

Combined with Equation (5.6) gives the equality claimed.

Using Lemmas 5.1 and 5.2, we proceed to the prove Theorem 3.1.

Proof of Theorem 3.1. Using Taylor expansion, for any $j \neq k$, we have

$$\sin \left( \frac{\pi}{2} \hat{\tau}_{jk} \right) - \sin \left( \frac{\pi}{2} \tau_{jk} \right) = \cos \left( \frac{\pi}{2} \tau_{jk} \right) \frac{\pi}{2} (\hat{\tau}_{jk} - \tau_{jk}) - \frac{1}{2} \sin(\theta_{jk}) \left( \frac{\pi}{2} \right)^2 (\hat{\tau}_{jk} - \tau_{jk})^2,$$

where $\theta_{jk}$ lies between $\tau_{jk}$ and $\hat{\tau}_{jk}$. Thus

$$\hat{\Sigma} - \Sigma = E_1 + E_2,$$
where $E_1, E_2 \in \mathbb{R}^{d \times d}$ satisfy that for $j \neq k$,
\[
\begin{align*}
[E_1]_{jk} &= \cos \left( \frac{\pi}{2} \tau_{jk} \right) \frac{\pi}{2} (\tilde{\tau}_{jk} - \tau_{jk}), \\
[E_2]_{jk} &= -\frac{1}{2} \sin(\theta_{jk}) \left( \frac{\pi}{2} \right)^2 (\tilde{\tau}_{jk} - \tau_{jk})^2,
\end{align*}
\]
and the diagonal entries of both $E_1$ and $E_2$ are all zero.

Using the results of U-statistics shown in Hoeffding (1963), we have that for any $j \neq k$ and $t > 0$,
\[
P(\|\tilde{\tau}_{jk} - \tau_{jk}\| > t) < 2e^{-nt^2/4}.
\]

For some constant $\alpha$, let the set $\Omega_2$ be defined as
\[
\Omega_2 := \left\{ \exists 1 \leq j \neq k \leq d, \|E_2\|_{jk} > \pi^2 \cdot \frac{\log(d/\alpha)}{n} \right\}.
\]
Since $\|E_2\|_{jk} \leq \frac{\pi^2}{8} (\tilde{\tau}_{jk} - \tau_{jk})^2$, by union bound, we have
\[
P(\Omega_2) \leq \frac{d^2}{2} \cdot 2e^{-2\log(d/\alpha)} = \alpha^2.
\]

Conditioning on $\Omega_2$, for any $v \in \mathbb{S}^{d-1}$, we have,
\[
|v^T E_2 v| \leq \sqrt{\sum_{j,k \in J} \|E_2\|_{jk}^2 \cdot \|v\|_2^2}
\]
\[
\leq \sqrt{d^2 \left( \pi^2 \cdot \frac{\log(d/\alpha)}{n} \right)^2}
\]
\[
= \pi^2 \cdot \frac{d \log(d/\alpha)}{n}. \quad (5.8)
\]

We then analyze the term $E_1$. Let $W = [W_{jk}] \in \mathbb{R}^{d \times d}$ with $W_{jk} = \frac{\pi}{2} \cos(\frac{\pi}{2} \tau_{jk})$ and $\hat{T} = [\hat{T}_{jk}]$ be the Kendall’s tau correlation matrix with $\hat{T}_{jk} = \tilde{\tau}_{jk}$. We can write
\[
E_1 = W \circ (\hat{T} - T),
\]
where $\circ$ represents the Hadamard product. Given the spectral norm bound of $\hat{T} - T$ shown in Lemma 5.1, we now focus on controlling $E_1$. Let $Y := (Y_1, \ldots, Y_d)^T \sim N_d(0, \Sigma)$ follows a Gaussian distribution with mean zero and covariance matrix $\Sigma$. Using the equality in Lemma 5.2, we have, for any $j \neq k$,
\[
E|Y_j Y_k| = \tau_{jk} \Sigma_{jk} + \frac{2}{\pi} \sqrt{1 - \Sigma_{jk}^2}.
\]
Reminding that
\[
\cos \left( \frac{\pi}{2} \tau_{jk} \right) = \sqrt{1 - \sin^2 \left( \frac{\pi}{2} \tau_{jk} \right)} = \sqrt{1 - \Sigma_{jk}^2},
\]
15
we have
\[ W_{jk} = \frac{\pi}{2} \cos \left( \frac{\pi}{2} \tau_{jk} \right) = \frac{\pi^2}{4} (E|Y_j Y_k| - \tau_{jk} \Sigma_{jk}). \]

Then let \( Y' := (Y'_1, \ldots, Y'_d)^T \in \mathbb{R}^d \) be an independent copy of \( Y \). We have, for any \( v \in \mathbb{S}^{d-1} \) and symmetric matrix \( M \in \mathbb{R}^{d \times d} \),

\[ |v^T M \circ W v| = \left| \sum_{j,k=1}^{d} v_j v_k M_{jk} W_{jk} \right| \]

\[ = \left| \frac{\pi^2}{4} \sum_{j,k} v_j v_k M_{jk} (|Y_j Y_k| - Y_j Y_k \text{sign}(Y'_j Y'_k)) \right| \]

\[ \leq \frac{\pi^2}{4} \mathbb{E} \left( \sum_{j,k} v_j v_k M_{jk} |Y_j Y_k| + \sum_{j,k} v_j v_k M_{jk} Y_j Y_k \text{sign}(Y'_j Y'_k) \right) \]

\[ \leq \frac{\pi^2}{4} \|M\|_2 \cdot \mathbb{E} \left( 2 \sum_j v_j^2 Y_j^2 \right) \]

\[ = \frac{\pi^2}{4} \|M\|_2 \cdot \left( 2 \sum_j v_j^2 \right) \]

\[ = \frac{\pi^2}{2} \|M\|_2. \quad (5.9) \]

Here the second inequality is due to the fact that for any \( M \in \mathbb{R}^{d \times d} \) and \( v \in \mathbb{R}^d \), \( |v^T M v| \leq \|M\|_2 \|v\|_2 \) and the third equality is due to the fact that \( \mathbb{E} Y_j^2 = \Sigma_{jj} = 1 \) for any \( j \in \{1, \ldots, d\} \). Accordingly, we have

\[ \|E_1\|_2 = \|W \circ (\hat{T} - T)\|_2 \leq \frac{\pi^2}{2} \|\hat{T} - T\|_2. \quad (5.10) \]

The bound in Theorem 3.1 follows from the fact that

\[ \|\hat{\Sigma} - \Sigma\|_2 = \|E_1 + E_2\| \leq \|E_1\|_2 + \|E_2\|_2 \]

and by combining Equations (5.1), (5.8), and (5.10). \( \square \)

5.2 Proofs of Propositions 4.5 and 4.6

Proposition 4.5 is a direct consequence of Lemma A.1. To prove Proposition 4.6, we first introduce the subexponential norm. For any random variable \( X \in \mathbb{R} \), \( \|X\|_{\phi_1} \) is defined as follows:

\[ \|X\|_{\phi_1} := \sup_{k \geq 1} \frac{1}{k} \left( \mathbb{E}|X|^k \right)^{1/k}. \]
Let \( S := \text{sign}(X - \tilde{X}) \). Because \( v^T S \) is subgaussian and \( \mathbb{E}v^T S = 0 \), using Lemma 5.14 in Vershynin (2010), we get
\[
\| (v^T S)^2 - \mathbb{E}(v^T S)^2 \|_{\phi_1} \leq \| (v^T S)^2 \|_{\phi_1} + \| v^T T v \|_{\phi_1}
\leq 2 \| v^T S \|_{\phi_2}^2 + v^T T v
\leq L_2 \| \Sigma \|_2 + \| T \|_2
\leq \max(L_2, 1)(\| \Sigma \|_2 + \| T \|_2).
\]

Since \( (v^T S)^2 - \mathbb{E}(v^T S)^2 \) is a zero-mean random variable and \( v^T S \) is subgaussian distributed, using Lemma 5.15 in Vershynin (2010), there exist two fixed constants \( C', c' \) such that if \( |t| \leq c'/(\| v^T S \|^2 - \mathbb{E}(v^T S)^2 \|_{\phi_1}^2) \), we have
\[
\mathbb{E}\exp(t((v^T S)^2 - \mathbb{E}(v^T S)^2)) \leq \exp(C't^2((v^T S)^2 - \mathbb{E}(v^T S)^2)^2_{\phi_1}).
\]

Accordingly, by choosing \( t_0 = c' \min(L_2^{-1}, 1)(\| \Sigma \|_2 + \| T \|_2)^{-1} \) and \( K = C' \max(L_2^2, 1) \) in Equation (4.3), noting that \( t_0 K((\| \Sigma \|_2 + \| T \|_2)^2 = c'C' \min(L_2^{-1}, 1) \max(L_2^2, 1)(\| \Sigma \|_2 + \| T \|_2) \geq 2c'C' \min(L_2, 1) \), the sign subgaussian condition is satisfied.

### 5.3 Proof of Theorem 4.7

In this section, we provide the proof of Theorem 4.7. In detail, we show that for any transelliptically distributed random vector \( X \) such that \( f(X) \sim N_d(0, I_d) \), we have that \( X \) satisfies the condition in Equation (4.3).

**Proof.** Because the sign function is invariant to strictly increasing functions, we only need to consider the random vector \( X \sim N_d(0, I_d) \). For \( X = (X_1, \ldots, X_d)^T \sim N_d(0, I_d) \) and \( \tilde{X} \) as an independent copy of \( X \), we have \( X - \tilde{X} \sim N_d(0, 2I_d) \). Reminding that the off-diagonal entries of \( I_d \) are all zero, defining \( X^0 = (X^0_1, \ldots, X^0_d)^T = X - \tilde{X} \) and
\[
g(X^0, v) := \sum_{j,k} v_j v_k \text{sign}(X^0_j X^0_k),
\]
we have
\[
\{v^T \text{sign}(X - \tilde{X})\}^2 - \mathbb{E}\{v^T \text{sign}(X - \tilde{X})\}^2 = g(X^0, v) - \mathbb{E}g(X^0, v).
\]

Accordingly, to bound \( \psi \left( \left\langle \text{sign}(X - \tilde{X}), v \right\rangle ; 2 \right) \), we only need to focus on \( g(X^0, v) \). Letting \( S := (S_1, \ldots, S_d)^T \) with \( S_j := \text{sign}(Y^0_j) \) for \( j = 1, \ldots, d \). Using the property of Gaussian distribution, \( S_1, \ldots, S_d \) are independent Bernoulli random variables in \( \{-1, 1\} \) almost surely. We then have
\[
g(Y^0, v) - \mathbb{E}g(Y^0, v) = \sum_{j,k} v_j v_k \text{sign}(Y^0_j Y^0_k) - 1 = (v^T S)^2 - 1.
\]
Here the first equality is due to the fact that $\|v\|_2^2 = \sum_{j=1}^d v_j^2 = 1$.

We then proceed to analyze the property of $(v^T S)^2 - 1$. By the Hubbard-Stratonovich transform (Hubbard, 1959), for any $\eta \in \mathbb{R}$,

$$\exp(\eta^2) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi}} e^{-y^2/4 + y\eta} dy. \tag{5.11}$$

Using Equation (5.11), we have that, for any $t > 0$,

$$\mathbb{E}\exp[t\{(v^T S)^2 - 1\}] = e^{-t\mathbb{E}e^{(v^T S)^2}}$$

$$= \frac{e^{-t}}{\sqrt{4\pi t}} \int_{-\infty}^{+\infty} e^{-y^2/4t} \sum_{j=1}^d v_j S_j dy$$

$$= \frac{e^{-t}}{\sqrt{4\pi t}} \int_{-\infty}^{+\infty} e^{-y^2/4t} \prod_{j=1}^d \frac{1}{2} (e^{yv_j} + e^{-yv_j}) dy.$$

For any number $z \in \mathbb{N}$, we define $z!$ represents the factorial product of $z$. Because for any $a \in \mathbb{R}$, by Taylor expansion, we have

$$\{\exp(a) + \exp(-a)\}/2 = \sum_{k=0}^{\infty} a^{2k}/(2k)!, \quad \text{and} \quad \exp(a^2/2) = \sum_{k=0}^{\infty} a^{2k}/(2k \cdot k!).$$

Because $(2k)! > 2^k \cdot k!$, we have

$$\{\exp(a) + \exp(-a)\}/2 \leq \exp(a^2/2).$$

Accordingly, we have for any $0 < t < 1/4$,

$$\mathbb{E}\exp[t\{(v^T S)^2 - 1\}] \leq \frac{e^{-t}}{\sqrt{1 - 2t}} \int_{-\infty}^{+\infty} e^{-y^2/(4t)} \sum_{j=1}^d \frac{1}{2} (e^{yv_j} + e^{-yv_j}) dy$$

$$\leq \frac{e^{-t}}{\sqrt{1 - 2t}} \int_{-\infty}^{+\infty} e^{-y^2/(4t)} \prod_{j=1}^d \frac{1}{2} (e^{yv_j} + e^{-yv_j}) dy$$

$$= \frac{e^{-t}}{\sqrt{1 - 2t}} \int_{-\infty}^{+\infty} e^{-y^2/(4t) + \frac{1}{2} y^2} dy$$

$$= \frac{e^{-t}}{\sqrt{1 - 2t}} \exp\left\{1 \sum_{k=1}^{\infty} \frac{(2t)^k}{k} \right\},$$

which implies that for all $0 < t < 1/4$,

$$\frac{e^{-t}}{\sqrt{1 - 2t}} \leq \exp\left(t^2 + \frac{1}{2} \sum_{k=3}^{\infty} \frac{(2t)^k}{k} \right) \leq \exp(2t^2).$$
This concludes that for $0 < t < 1/4$,
\[
\mathbb{E}\exp[t\{(v^T S)^2 - 1\}] \leq \exp(2t^2). \tag{5.12}
\]
Due to that $(v^T S)^2 \geq 0$, we can apply Theorem 2.6 in Chung and Lu (2006) to control the term $\mathbb{E}\exp[t\{1 - (v^T S)^2\}]$. In detail, suppose that the random variable $Y$ satisfying $\mathbb{E}Y = 0$, $Y \leq a_0$, and $\mathbb{E}Y^2 = b_0$ for some absolute constants $a_0$ and $b_0$. Then for any $0 < t < 2/a_0$, using the proof of Theorem 2.8 in Chung and Lu (2006), we have,
\[
\mathbb{E}e^{tY} \leq \exp\{3b_0/2 \cdot t^2\}. \tag{5.13}
\]
For $Y = 1 - (v^T S)^2$, we have
\[
a_0 = 1 \quad \text{and} \quad b_0 = \mathbb{E}(v^T S)^4 - 1 = 2 - 2 \sum_{j=1}^{d} v_j^4 < 2. \tag{5.14}
\]
Here we remind that $\mathbb{E}(v^T S)^2 = \sum_j v_j^2 = 1$. Combining Equations (5.13) and (5.14) implies that for any $t > 0$,
\[
\mathbb{E}\exp[t\{1 - (v^T S)^2\}] \leq \exp\{3t^2\}. \tag{5.15}
\]
Combining Equations (5.12) and (5.15), we see that Equation (4.3) holds with $K = 3/4$ and $t_0 = 1/4$ (Reminding that here $\|\Sigma\|_2 = \|T\|_2 = 1$).

### 5.4 Proof of Theorem 4.8 and Corollary 4.9

In this section, we prove Theorem 4.8 and Corollary 4.9. Using the same argument as in the proof of Theorem 4.7, because the sign function is invariant to the marginal strictly increasing transformation function, we only need to focus on those random vectors that are Gaussian distributed.

**Proof of Theorem 4.8.** Assume that $\Sigma \in \mathbb{R}^{d \times d}$ is a compound symmetry matrix such that
\[
\Sigma_{jj} = 1 \quad \text{and} \quad \Sigma_{jk} = \rho \quad \text{for} \ j \neq k.
\]
By the discussion in Page 11 of Vershynin (2010), to prove Equation (4.5) holds, we only need to prove that for $0 \leq \rho \leq C_0$ where $C_0$ is some absolute constant, $X = (X_1, \ldots, X_d)^T \sim N_d(0, \Sigma)$ and $v \in \mathbb{S}^{d-1}$, we have
\[
\exp(tv^T \text{sign}(X - \bar{X})) \leq \exp(c\|\Sigma\|_2 t^2),
\]
for some fixed constant $c$. This result can be proved as follows. Let $\eta_0, \eta_1, \ldots, \eta_d$ be i.i.d. standard Gaussian distribution such that $\eta_i \sim N_1(0,1)$ for $i = 0, 1, \ldots, d$, we then have
\[ Z := X - \tilde{X} \] can be expressed as \( Z \overset{d}{=} (Z'_1, \ldots, Z'_d)^T \), where
\[
Z'_1 = \sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_1}},
Z'_2 = \sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_2}},
\]
\[
\ldots,
Z'_d = \sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_d}}.
\]

Accordingly, we have
\[
\mathbb{E} \exp(tv^T \text{sign}(X - \tilde{X})) = \mathbb{E}\left( \exp\left( t \sum_{j=1}^{d} v_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}}) \right) \right)
= \mathbb{E}\left( \mathbb{E}(\exp(t \sum_{j=1}^{d} v_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j})|\eta_0)) \right)
\]

Moreover, we have
\[
\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}}|\eta_0 \sim N_1(\sqrt{2\rho \eta_0, 2 - 2\rho}). \tag{5.16}
\]

Letting \( \mu := \sqrt{2\rho \eta_0} \) and \( \sigma := \sqrt{2 - 2\rho} \), Equation (5.16) implies that
\[
\mathbb{P}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}} > 0|\eta_0) = \Phi\left( \frac{\mu}{\sigma} \right),
\]
where \( \Phi(\cdot) \) is the CDF of the standard Gaussian. This further implies that
\[
\text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}})|\eta_0 \sim \text{Bern}\left( \Phi\left( \frac{\mu}{\sigma} \right) \right),
\]
where we denote \( Y \sim \text{Bern}(p) \) if \( \mathbb{P}(Y = 1) = p \) and \( \mathbb{P}(Y = -1) = 1 - p \). Accordingly, letting \( \alpha := \Phi(\mu/\sigma) \), we have
\[
\mathbb{E}\left( \exp(tv_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j})|\eta_0) = (1 - \alpha)e^{-v_j t} + \alpha e^{v_j t}.
\]

Letting \( \beta := \alpha - 1/2 \), we have
\[
\mathbb{E}\left( \exp(tv_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j})|\eta_0) = \frac{1}{2} e^{-v_j t} + \frac{1}{2} e^{v_j t} + \beta(e^{v_j t} - e^{-v_j t}).
\]

Using that fact that \( \frac{1}{2} e^{a} + \frac{1}{2} e^{-a} \leq e^{a^2/2}, \) we have
\[
\mathbb{E}\left( \exp(tv_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j})|\eta_0) \leq \exp(v_j^2 t^2/2) + \beta(e^{v_j t} - e^{-v_j t}).
\]

Because conditioning on \( \eta_0 \), \( \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}) \) are independent of each other, we have
\[
\mathbb{E}\left( \exp\left( t \sum_{j=1}^{d} v_j \text{sign}(\sqrt{2\rho \eta_0 + \sqrt{2 - 2\rho \eta_j}) \right) |\eta_0) \leq \prod_{j=1}^{d} \left\{ \exp(v_j^2 t^2/2) + \beta(e^{v_j t} - e^{-v_j t}) \right\}
= e^{t^2/2} \left( 1 + \sum_{k=1}^{d} \beta^k \sum_{j_1 < j_2 < \cdots < j_k} \prod_{j \in \{j_1, \ldots, j_k\}} \frac{e^{v_j t} - e^{-v_j t}}{e^{v_j^2 t^2/2}} \right).
\]
Moreover, for any centered Gaussian distribution $Y \sim N_1(0, \kappa)$ and $t \in \mathbb{R}$, we have

$$\mathbb{P}(\Phi(Y) > 1/2 + t) = \mathbb{P}(Y > \Phi^{-1}(1/2 + t)) = \mathbb{P}(Y > -\Phi^{-1}(1/2 - t))$$

$$= \mathbb{P}(Y < -\Phi^{-1}(1/2 - t))$$

$$= \mathbb{P}(\Phi(Y) < 1/2 - t).$$

Combined with the fact that $\Phi(Y) \in [0, 1]$, we have

$$\mathbb{E}(\Phi(Y) - 1/2)^k = 0 \quad \text{when } k \text{ is odd.}$$

This implies that when $k$ is odd,

$$\mathbb{E}\beta^k = 0 = \mathbb{E}\left(\Phi(\sqrt{\rho/(1 - \rho)}\eta_0) - \frac{1}{2}\right)^k = 0.$$

Accordingly, denoting $\epsilon = \mathbb{E}\exp\left(t \sum_{j=1}^d v_j \text{sign}(\sqrt{2\rho}\eta_0 + \sqrt{2 - 2\rho}j)\right)$, we have

$$\epsilon \leq e^{t^2/2} \left(1 + \sum_{k \text{ is even}} \mathbb{E}\beta^k \sum_{j_1 < j_2 < \cdots < j_k} \prod_{j \in \{j_1, \ldots, j_k\}} \frac{e^{v_j t} - e^{-v_j t}}{e^{v_j^2 t^2/2}}\right).$$

Using the fact that

$$|e^a - e^{-a}| = \left|\sum_{j=1}^{\infty} \frac{a^j}{j!} - \sum_{j=1}^{\infty} \frac{(-a)^j}{j!}\right|$$

$$= 2 \left|\sum_{m=0}^{\infty} \frac{a^{2m+1}}{(2m+1)!}\right|$$

$$= 2|a| \cdot \left|\sum_{m=0}^{\infty} \frac{a^{2m}}{(2m+1)!}\right|$$

$$\leq 2|a| \exp(a^2/2),$$

we further have

$$\epsilon \leq e^{t^2/2} \left(1 + \sum_{k \text{ is even}} \mathbb{E}\beta^k \sum_{j_1 < j_2 < \cdots < j_k} \prod_{j \in \{j_1, \ldots, j_k\}} 2|v_j|\right)$$

$$= e^{t^2/2} \left(1 + \sum_{k \text{ is even}} \mathbb{E}\beta^k (2|t|)^k \sum_{j_1 < j_2 < \cdots < j_k} |v_{j_1} \cdots v_{j_k}|\right).$$

By Maclaurin’s inequality, for any $x_1, \ldots, x_d \geq 0$, we have

$$\frac{x_1 + \cdots + x_n}{n} \geq \left(\frac{\sum_{1 \leq i < j \leq n} x_i x_j}{\binom{n}{2}}\right)^{1/2} \geq \cdots \geq (x_1 \cdots x_n)^{1/n}.$$
Accordingly, 

\[ e^{t^2/2} \left( 1 + \sum_{k \text{ is even}} E \beta^k (2|t|)^k \sum_{j_1 < j_2 < \cdots < j_k} |v_{j_1} \cdots v_{j_k}| \right) \]

\[ \leq e^{t^2/2} \left( 1 + \sum_{k \text{ is even}} E \beta^k (2|t|)^k \{ \|v\|_1/d \} \right) \]

\[ \leq e^{t^2/2} \left( 1 + \sum_{k \text{ is even}} E \beta^k (2|t|)^k d^{k/2} (e/k)^k \right). \] (5.17)

The last inequality is due to the fact that \( \|v\|_1 \leq \sqrt{d}\|v\|_2 = \sqrt{d} \) and \( \{d\} \leq (ed/k)^k \).

Finally, we analyze \( E \beta^{2m} \) for \( m = 1, 2, \ldots \). Reminding that

\[ \beta := \Phi \left( \sqrt{\frac{\rho}{1-\rho}} \eta_0 \right) - \frac{1}{2}, \]

consider the function \( f(x) : x \to \Phi(\sqrt{\rho/(1-\rho)}x) \), we have

\[ |f'(x)| = \sqrt{\frac{\rho}{1-\rho}} \cdot \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{\rho}{2(1-\rho)} x^2 \right) \leq \sqrt{\frac{\rho}{2\pi(1-\rho)}}. \]

Accordingly, \( f(\cdot) \) is a Lipschitz function with a Lipschitz constant \( K_0 := \sqrt{\frac{\rho}{2\pi(1-\rho)}} \). By the concentration of Lipschitz functions of Gaussian (Ledoux, 2001), we have

\[ P(|\beta| > t) = P(|f(\eta_0) - E f(\eta_0)| > t) \leq 2 \exp(-t^2/(2K_0^2)). \]

This implies that, for \( m = 1, 2, \ldots \),

\[ E \beta^{2m} = 2m \int_0^\infty t^{2m-1} P(|\beta| > t) dt \]

\[ \leq 4m \int_0^\infty t^{2m-1} \exp\left(-t^2/(2K_0^2)\right) dt \]

\[ = 4m(\sqrt{2K_0})^{2m} \int_0^\infty t^{2m-1} \exp(-t^2) dt \]

\[ = 2m(2K_0^2)^m \int_0^\infty t^{m-1} \exp(-t) dt. \]

Using the fact that \( \int_0^\infty \exp(-t) dt = 1 \) and for any \( m \geq 1 \),

\[ m \int_0^\infty t^{m-1} \exp(-t) dt = \int_0^\infty \exp(-t) dt^m = \int_0^\infty t^m \exp(-t) dt, \]

we have for \( m \in \mathbb{Z}^+, \int_0^\infty t^m \exp(-t) dt = m! \). Accordingly,

\[ E \beta^{2m} \leq 2m(2K_0^2)^m (m-1)! = 2(2K_0^2)^m m!. \]
Plugging the above result into Equation (5.17), we have

\[ \epsilon \leq e^{t^2/2} \left( 1 + \sum_{m=1}^{\infty} 2(2K_0^2)^m m! (2t)^{2m} d^m (e/(2m))^{2m} \right) \]

\[ = e^{t^2/2} \left( 1 + \sum_{m=1}^{\infty} (K_0^2 d)^m m! 2(2\sqrt{2} et)^{2m}/(2m)^{2m} \right). \]

Reminding that \( \rho \leq C_0 \) and \( K_0 := \sqrt{\frac{\rho}{2\pi(1-\rho)}} \leq \sqrt{\frac{\rho}{2\pi(1-C_0)}} \), we have

\[ \epsilon \leq e^{t^2/2} \left( 1 + \sum_{m=1}^{\infty} (K_0^2 d)^m m! 2(2\sqrt{2} et)^{2m}/(2m)^{2m} \right) \]

\[ \leq e^{t^2/2} \left( 1 + \sum_{m=1}^{\infty} m! 2 \left( 2\sqrt{\frac{d\rho}{\pi(1-C_0)}} t \right)^{2m}/(2m)^{2m} \right). \]

Finally, we have for any \( m \geq 1 \)

\[ 2m! \cdot m! \leq (2m)^{2m}, \]

implying that

\[ \epsilon \leq e^{t^2/2} \cdot \exp(4d\rho e^2/\pi \cdot t^2) = \exp \left\{ \left( \frac{1}{2} + \frac{4d\rho e^2}{\pi(1-C_0)} \right) t^2 \right\}, \quad (5.18) \]

where the term \( \frac{1}{2} + \frac{4d\rho e^2}{\pi(1-C_0)} \) is in the same scale of \( \|\Sigma\|_2 = 1 + (d-1)\rho \). This finishes the proof.

Corollary 4.9 can be proved similarly as Theorem 4.8.

**Proof of Corollary 4.9.** Letting \( J_k = \{1 + \sum_{j=1}^{k-1} d_j, \ldots, \sum_{j=1}^{k} d_j\} \). By independence of the Gaussian distribution, we have

\[ \exp(tv^T \text{sign}(X - \tilde{X})) = \prod_{k=1}^{q} \exp(tv^T_{J_k} \text{sign}(X - \tilde{X})_{J_k}). \]

Here we note that the bound in Equation (5.18) also holds for each \( \exp(tv^T_{J_k} \text{sign}(X - \tilde{X})_{J_k}) \) by checking Equation (5.17). Accordingly,

\[ \prod_{k=1}^{q} \exp(tv^T_{J_k} \text{sign}(X - \tilde{X})_{J_k}) \leq \prod_{k=1}^{q} \exp \left\{ \left( \frac{1}{2} + \frac{4d_k\rho_k e^2}{\pi(1-C_1)} \right) t^2 \right\} \]

\[ \leq \exp \left\{ t^2 \left( \frac{q}{2} + \frac{4e^2 q}{\pi(1-C_1)} \max_k (d_k\rho_k) \right) \right\}, \]

Because \( q \) is upper bounded by a fixed constant, we have \( v^T \text{sign}(X - \tilde{X}) \) is a subgaussian distribution. This finishes proof. \( \square \)
5.5 Proof of Theorem 4.10

To prove the result in Theorem 4.10, we first introduce the following lemma, which quantifies the distance between $\hat{\theta}_1$ and $\theta_1$.

**Lemma 5.3** (Vu and Lei (2012)). When the conditions in Theorem 4.10 hold, we have

$$\sin \angle(\hat{\theta}_1, \theta_1) \leq \frac{2}{\lambda_1(\Sigma) - \lambda_2(\Sigma)} \sup_{v \in \mathbb{S}^{d-1}, \|v\|_0 \leq 2s} |v^T(\hat{\Sigma} - \Sigma)v|.$$ 

With Lemma 5.3, we now proceed to the next lemma, which shows that under the conditions in Theorem 4.10, $\sup_{v \in \mathbb{S}^{d-1}, \|v\|_0 \leq 2s} |v^T(\hat{T} - T)v| = O_P(\sqrt{s \log d/n})$. Here we remind that $T$ and $\hat{T}$ are the population and empirical Kendall’s tau matrix, with $T_{jk} = \tau_{jk}$ and $\hat{T}_{jk} = \hat{\tau}_{jk}$. We denote $J_q$ to be any subset of $\{1, \ldots, d\}$ with the cardinality $q$.

**Lemma 5.4.** Under the conditions in Theorem 4.10, letting $\zeta := K(\|\Sigma\|_2 + \|T\|_2)^2$, we have with probability larger than or equal to $1 - 2\alpha$,

$$\sup_{b \in \mathbb{S}^{2s-1}} \sup_{J_{2s} \subseteq \{1, \ldots, d\}} \left| b^T[\hat{T} - T]_{J_{2s} \times J_{2s}} b \right| \leq 2(8\zeta)^{1/2} \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}}. \quad (5.19)$$

**Proof.** Let $q \in \mathbb{Z}^+$ be larger than 1. For any sphere $\mathbb{S}^{q-1}$ equipped with Euclidean distance, we let $N_\epsilon$ be a subset of $\mathbb{S}^{q-1}$ such that for any $v \in \mathbb{S}^{q-1}$, there exists $u \in N_\epsilon$ subject to $\|u - v\|_2 \leq \epsilon$. Let $N_{1/4}$ be a $(1/4)$-net of $\mathbb{S}^{2s-1}$. The cardinal number has the upper bound

$$\text{card}(N_\epsilon) < \left(1 + \frac{2}{\epsilon}\right)^{2s},$$

thus bounded by $9^{2s}$. Moreover, for any matrix $M \in \mathbb{R}^{2s \times 2s}$,

$$\sup_{v \in \mathbb{S}^{2s-1}} |v^T M v| \leq \frac{1}{1 - 2\epsilon} \sup_{v \in N_\epsilon} |v^T M v|.$$ 

This implies that

$$\sup_{v \in \mathbb{S}^{2s-1}} |v^T M v| \leq 2 \sup_{v \in N_{1/4}} |v^T M v|.$$ 

Let $\beta > 0$ be a constant defined as

$$\beta := (8\zeta)^{1/2} \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}}.$$
We have
\[
\mathbb{P}\left( \sup_{b \in S^{2s-1}} \sup_{J_{2s} \subset \{1, \ldots, d\}} \left| b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} \right| > 2\beta \right) \\
\leq \mathbb{P}\left( \sup_{b \in \mathcal{N}_{1/4}} \sup_{J_{2s} \subset \{1, \ldots, d\}} \left| b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} \right| > \beta \right) \\
\leq 9^{2s} \left( \frac{d}{2s} \right) \mathbb{P}\left( \left| b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} \right| > (8\zeta)^{1/2} \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}}, \right.
\]
for fixed \(b\) and \(J_{2s}\).

Thus, if we can show that for any fixed \(b\) and \(J_{2s}\), we have
\[
\mathbb{P}\left( \left| b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} \right| > t \right) \leq 2e^{-nt^2/(8\zeta)},
\] (5.20)
for the absolute constant \(\zeta := K(\|\Sigma\|_2 + \|T\|_2)^2\). Then using the bound \(\left( \frac{d}{2s} \right) < \{ed/(2s)\}^{2s}\), we have
\[
9^{2s} \left( \frac{d}{2s} \right) \mathbb{P}\left( \left| b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} \right| > (8\zeta)^{1/2} \sqrt{\frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n}}, \right.
\]
for fixed \(b\) and \(J\)
\[
\leq 2 \exp\{2s(1 + \log 9 - \log(2s)) + 2s \log d - 2s(3 + \log d - \log(2s)) - \log(1/\alpha)\}
\leq 2\alpha,
\]
It gives that with probability greater than \(1 - 2\alpha\) the bound in Equation (5.19) holds.

Finally, we show that Equation (5.20) holds. For any \(t\), we have
\[
\mathbb{E} \exp \left\{ t \cdot b^T \left[ \hat{T} - T \right]_{J_{2s}, J_{2s}} b \right\}
\leq \mathbb{E} \exp \left\{ t \cdot \sum_{j \neq k \in J_{2s}} b_j b_k (\hat{\tau}_{jk} - \tau_{jk}) \right\}
\leq \mathbb{E} \exp \left\{ t \cdot \sum_{i < i'} \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}(x_i - x_{i'})(x_i - x_{i'})_j - \tau_{jk}) \right\}.
\]
Let \(S_n\) represent the permutation group of \(\{1, \ldots, n\}\). For any \(\sigma \in S_n\), let \((i_1, \ldots, i_n) := \sigma(1, \ldots, n)\) represent a permuted series of \(\{1, \ldots, n\}\) and \(O(\sigma) := \{(i_1, i_2), (i_3, i_4), \ldots, (i_{n-1}, i_n)\}\).
In particular, we denote $O(\sigma_0) := \{(1, 2), (3, 4), \ldots, (n-1, n)\}$. By simple calculation,

$$E \exp \left\{ t \cdot \frac{1}{2} \sum_{i < i'} \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((x_i - x_{i'})_j (x_i - x_{i'})_k) - \tau_{jk}) \right\}$$

$$= E \exp \left\{ t \cdot \frac{1}{\text{card}(S_n)} \sum_{\sigma \in S_n} \frac{2}{n} \sum_{(i, i') \in \sigma} \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((x_i - x_{i'})_j (x_i - x_{i'})_k) - \tau_{jk}) \right\}$$

$$\leq \frac{1}{\text{card}(S_n)} \sum_{\sigma \in S_n} E \exp \left\{ t \cdot \frac{2}{n} \sum_{(i, i') \in \sigma} \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((x_i - x_{i'})_j (x_i - x_{i'})_k) - \tau_{jk}) \right\}$$

$$= E \exp \left\{ t \cdot \frac{2}{n} \sum_{(i, i') \in \sigma} \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((x_i - x_{i'})_j (x_i - x_{i'})_k) - \tau_{jk}) \right\}.$$  \hspace{1cm} (5.21)

The inequality is due to the Jensen's inequality.

Let $m := n/2$ and remind that $X = (X_1, \ldots, X_d)^T \sim T E_d(\Sigma, \xi; f_1, \ldots, f_d)$. Let $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_d)^T$ be an independent copy of $X$. By Equation (4.3), we have that for any $|t| < t_0$ and $v \in S^{d-1},$

$$E \exp[t\{ (v^T \text{sign}(X - \overline{X}))^2 - E(v^T \text{sign}(X - \overline{X}))^2 \}] \leq e^{|t|^2}.$$

In particular, letting $v_{J_{2s}} = b$ and $v_{J_{2s}^C} = 0$, we have

$$E \exp \left\{ t \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((X - \overline{X})_j (X - \overline{X})_k) - \tau_{jk}) \right\} \leq e^{|t|^2}. \hspace{1cm} (5.22)$$

Then we are able to continue Equation (5.21) as

$$E \exp \left\{ t \cdot \frac{2}{m} \sum_{i = 1}^m \sum_{j \neq k \in J_{2s}} b_j b_k (\text{sign}((x_{2i} - x_{2i-1})_j (x_{2i} - x_{2i-1})_k) - \tau_{jk}) \right\}$$

$$= \left( E e^{t/2 (\text{sign}((X - \overline{X})_j (X - \overline{X})_k) - \tau_{jk})} \right)^m \leq e^{|t|^2/m}, \hspace{1cm} (5.23)$$

where by Equation (4.3), the last inequality holds for any $|t/m| < t_0$. Accordingly, choosing $t = \beta m/(2\zeta)$, by Markov inequality, we have

$$P \left( b^T [\overline{T} - T]_{J_{2s}^C, J_{2s}} b > \beta \right) \leq e^{-n\beta^2/(8\zeta)}, \text{ for all } \beta < 2\zeta t_0. \hspace{1cm} (5.24)$$

By symmetry, we have the same bound for $P \left( b^T [\overline{T} - T]_{J_{2s}, J_{2s}^C} b < -\beta \right)$ as in Equation (5.24). Together they give us Equation (5.20). This completes the proof.\hfill\Box
Using the result in Lemma 5.4, we can now proceed to the next lemma, which quantifies the term \( \sup_{v \in \mathbb{S}^{d-1}, \|v\|_0 \leq 2s} |v^T(\hat{\Sigma} - \Sigma)v| \).

**Lemma 5.5.** Under the conditions in Theorem 4.10, we have with probability larger than or equal to \( 1 - 2\alpha - \alpha^2 \),

\[
\sup_{b \in \mathbb{S}^{2s-1}} \sup_{J_{2s} \in \{1, \ldots, d\}} \left| b^T [\hat{\Sigma} - \Sigma]|_{J_{2s},J_{2s}} b \right| 
\leq \pi^2 \left( 8\zeta \right)^{1/2} \left( \frac{2s(3 + \log(d/2s)) + \log(1/\alpha)}{n} \right) + 2\pi^2 \cdot \frac{s \log(d/\alpha)}{n}. \tag{5.25}
\]

**Proof.** Using a similar argument as in the proof of Theorem 3.1, we let \( \mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{d \times d} \), satisfying that for \( j \neq k \),

\[
[\mathbf{E}_1]_{jk} = \cos \left( \frac{\pi}{2} \tau_{jk} \right) \frac{\pi}{2} (\hat{\tau}_{jk} - \tau_{jk}),
\]

\[
[\mathbf{E}_2]_{jk} = -\frac{1}{2} \sin(\theta_{jk}) \left( \frac{\pi}{2} \right)^2 (\hat{\tau}_{jk} - \tau_{jk})^2.
\]

We then have

\[\hat{\Sigma} - \Sigma = \mathbf{E}_1 + \mathbf{E}_2.\]

Let the set \( \Omega_2 \) be defined as

\[\Omega_2 := \left\{ \exists 1 \leq j \neq k \leq d, \|\mathbf{E}_2|_{jk}\| > \frac{\pi^2 \log(d/\alpha)}{n} \right\}.\]

Using the result in the proof of Theorem 3.1, we have \( \mathbb{P}(\Omega_2) \leq \alpha^2 \). Moreover, conditioning on \( \Omega_2 \), for any \( J_{2s} \in \{1, \ldots, d\} \) and \( b \in \mathbb{S}^{2s-1} \),

\[
|b^T [\mathbf{E}_2]|_{J_{2s},J_{2s}} b| \leq \sqrt{\sum_{j,k \in J_{2s}} \|\mathbf{E}_2|_{jk}\| \cdot \|b\|_2^2}
\leq (2s) \cdot \frac{\pi^2 \cdot \log(d/\alpha)}{n}
= 2\pi^2 \cdot \frac{s \log(d/\alpha)}{n}. \tag{5.26}
\]

We then proceed to control the term \( |b^T [\mathbf{E}_1]|_{J_{2s},J_{2s}} b| \). Using a similar argument as shown in Equation (5.9), for \( \mathbf{Y} = (Y_1, \ldots, Y_d)^T \sim N_d(0, \Sigma) \), any symmetric matrix \( \mathbf{M} \in \mathbb{R}^{d \times d} \), \( \mathbf{W} \)
with \( W_{jk} = \frac{\pi}{2} \cos\left(\frac{\pi}{2} \tau_{jk}\right) \) and \( v \in S^{d-1} \) with \( \|v\|_0 \leq q \), we have
\[
|v^T M \circ Wv| \leq \frac{\pi^2}{4} \mathbb{E} \left( \left| \sum_{j,k} v_j v_k M_{jk} |Y_j Y_k| \right| + \left| \sum_{j,k} v_j v_k M_{jk} Y_j Y_k \text{sign}(Y_j Y_k) \right| \right)
\]
\[
\leq \frac{\pi^2}{4} \sup_{b \in S^{d-1}, \|b\|_0 \leq q} |b^T Mb| \cdot \mathbb{E} \left( 2 \sum_j v_j^2 Y_j^2 \right)
\]
\[
= \frac{\pi^2}{4} \sup_{b \in S^{d-1}, \|b\|_0 \leq q} |b^T Mb| \cdot \left( 2 \sum_j v_j^2 \right)
\]
\[
= \frac{\pi^2}{2} \sup_{b \in S^{d-1}, \|b\|_0 \leq q} |b^T Mb|.
\]

Accordingly, we have
\[
\sup_{b \in S^{2s-1}} \sup_{J_{2s} \in \{1, \ldots, d\}} \left| b^T [E_1]_{J_{2s},J_{2s}} b \right| \leq \frac{\pi^2}{2} \sup_{b \in S^{2s-1}} \sup_{J_{2s} \in \{1, \ldots, d\}} \left| b^T \left[ T - \hat{T} \right]_{J_{2s},J_{2s}} b \right|.
\]

Combined with Lemma 5.4 and Equation (5.26), we have the desired result. \( \square \)

*Proof of Theorem 4.10.* Combining Lemmas 5.3 and 5.5 provides the bound in Theorem 4.10. \( \square \)

### 6 Discussions

This paper considers the optimal rates of convergence of transelliptical component analysis. We mainly focus on estimating the leading eigenvector of the latent generalized correlation matrix of a transelliptical distribution. The main procedure is to exploit a rank-based method using the Kendall's tau correlation coefficient. Minimax rates of convergence are discussed under both non-sparse and sparse settings. Under the non-sparse settings, we show that a PCA version of TCA has been rate optimal. Under the non-sparse setting, we analyze the estimator proposed in Han and Liu (2012b) and provide a subfamily of the transelliptical distributions, under which this estimator attains the minimax rate of convergence. It should also be noted that unlike the Gaussian case, the theoretical analysis under the transelliptical family motivates new understandings on rank-based estimators as well as new proof techniques. These new understandings and proof techniques are of self interests.

Moreover, Han and Liu (2013) study the performance of TCA on dependent data under some mixing conditions and prove that TCA can attain a \( s \sqrt{\log d/(n\gamma)} \) rate of convergence, where \( \gamma \leq 1 \) reflects the impact of non-independence on the estimation accuracy. It is also interesting to consider extending the results in this paper to dependent data under similar...
mixing conditions and see whether a similar $\sqrt{s \log d/(n\gamma')} \infty$ rate of convergence can be attained. However, it is much more challenging to obtain such results in non-independent data. The current theoretical analysis based on U statistics is not sufficient to achieve this goal.

A problem closely related to the leading eigenvector estimation is principal component detection, which is initiated in the work of Berthet and Rigollet (2013b,a). It is interesting to extend TCA to this setting and conduct sparse principal component detection under the transelliptical family. It is worth pointing out that Theorems 3.1 and 4.10 in this paper can be exploited in measuring the statistical performance of the corresponding detection of sparse principal components.

**Acknowledgement**

We sincerely thank Marten Wegkamp for his very helpful discussions and generously providing independent credit for our work. We also thank Xiuyuan Cheng, Ramon van Handel, Philippe Rigollet, and Luo Xiao for their many helps that significantly improve the quality of our work.

**A Appendix**

In this section we provide a lemma quantifying the relationship between Orlicz $\psi_2$-norm and the subgaussian condition. Although this result is well known, in order to quantify this relationship in numbers, we include a proof here. We do not claim any original contribution in this section.

**Lemma A.1.** For any random variable $Y \in \mathbb{R}$, we say that $Y$ has a subgaussian distribution with factor $c > 0$ if and only if for any $t \in \mathbb{R}$, $\mathbb{E} \exp(tY) \leq \exp(ct^2)$. We than have $Y$ has a subgaussian distribution if and only if $\|Y\|_{\psi_2}$ is bounded. In particular, we have that if $Y$ is subgaussian with factor $c$, then $\|Y\|_{\psi_2} \leq \sqrt{12c}$. If $\|Y\|_{\psi_2} \leq D \leq \infty$, then $Y$ is subgaussian with factor $c = 5D^2/2$.

**Proof.** If $Y$ has a subgaussian distribution, then for any $m > 0$, we have

$$
\mathbb{E} \exp(|Y/m|^2) = 1 + \int_0^\infty \mathbb{P} \left( \frac{Y^2}{m^2} > t \right) e^t dt
= 1 + \int_0^{\infty} \mathbb{P}(|Y| > m \sqrt{t}) e^t dt.
$$

By Markov inequality, we know that if $Y$ is subgaussian distributed, then for any $t > 0$

$$
\mathbb{P}(|Y| > t) \leq 2 \exp(-t^2/(4c)).
$$
Accordingly, we can proceed the proof
\[
E \exp(|Y/m|^2) \leq 1 + 2 \int_0^\infty e^{-m^2t/(4c)} \cdot e^t dt
= 1 + 2 \int_0^\infty e^{-(m^2/(4c)-1)t} dt
= 1 + \frac{2}{m^2/(4c)-1}.
\]
Picking \( m = \sqrt{2c} \), we have \( E \exp(|Y/m|^2) \leq 2 \). Accordingly, \( \|Y\|_\psi \leq \sqrt{2c} \). On the other hand, if \( \|Y\|_\psi \leq \infty \), then there exists some \( m < \infty \) such that \( E \exp(|Y/m|^2) \leq 2 \). Using integration by part, it is easy to check that
\[
\exp(a) = 1 + a^2 \int_0^1 (1 - y)e^{ay} dy.
\]
This implies that
\[
E \exp(tX) = 1 + \int_0^1 (1 - u)E[(tX)^2 \exp(utX)]du
\leq 1 + t^2E(X^2 \exp(|tX|)) \int_0^1 (1 - u)du
\leq 1 + \frac{t^2}{2}E(X^2e^{tX}).
\]
Using the fact that for any \( a, b \in \mathbb{R}, |ab| \leq \frac{a^2+b^2}{2} \) and \( a \leq e^a \), we can further prove that
\[
E \exp(tX) \leq 1 + \frac{t^2}{2}E(X^2e^{tX})
\leq 1 + m^2t^2e^{m^2t^2/(2m^2)}E \left( \frac{X^2}{2m^2} e^{X^2/(2m^2)} \right)
\leq 1 + m^2t^2e^{m^2t^2/(2m^2)}e^{X^2/m^2}
\leq (1 + 2m^2t^2)e^{m^2t^2/2}
\leq e^{5m^2t^2/2}.
\]
The last inequality is due to the fact that for any \( a \in \mathbb{R}, 1 + a \leq e^a \). Accordingly, \( X \) is subgaussian with the factor \( c = 5m^2/2 \).

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, volume 2. Wiley.

Berthet, Q. and Rigollet, P. (2013a). Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828.*
Berthet, Q. and Rigollet, P. (2013b). Optimal detection of sparse principal components in high dimension. *to appear Annals of Statistics.*

Bunea, F. and Xiao, L. (2012). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *arXiv preprint arXiv:1212.5321.*

Cai, T. T., Ma, Z., and Wu, Y. (2012). Sparse pca: Optimal rates and adaptive estimation. *arXiv preprint arXiv:1211.1309.*

Chung, F. R. K. and Lu, L. (2006). *Complex graphs and networks.* Number 107. American Mathematical Society.

d’Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G. (2004). A direct formulation for sparse pca using semidefinite programming. *SIAM Review,* 49:434–448.

Fang, H., Fang, K., and Kotz, S. (2002). The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis,* 82:1–16.

Fang, K., Kotz, S., and Ng, K. (1990). *Symmetric Multivariate and Related Distributions.* Chapman&Hall.

Han, F. and Liu, H. (2012a). Semiparametric principal component analysis. In *Proceedings of the Twenty-fifth Annual Conference on Neural Information Processing Systems,* pages 171–179.

Han, F. and Liu, H. (2012b). Transelliptical component analysis. In *Proceedings of the Twenty-fifth Annual Conference on Neural Information Processing Systems,* pages 368–376.

Han, F. and Liu, H. (2013). Principal component analysis on non-gaussian dependent data. *Proceedings of the Thirtieth International Conference on Machine Learning.*

Han, F., Zhao, T., and Liu, H. (2013). Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research,* 14:629–671.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association,* 58:13–30.

Hubbard, J. (1959). Calculation of partition functions. *Physical Review Letters,* 3(2):77.

Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association,* 104:682–693.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics,* 12:531–547.
Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11:517–553.

Jung, S. and Marron, J. (2009). Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89. American Mathematical Society.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012a). High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40:2293–2326.

Liu, H., Han, F., and Zhang, C.-H. (2012b). Transelliptical graphical models. In *Proceedings of the Twenty-fifth Annual Conference on Neural Information Processing Systems*, pages 809–817.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328.

Lounici, K. (2012). High-dimensional covariance matrix estimation with missing observations. *arXiv preprint arXiv:1201.2577*.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *to appear Annals of Statistics*.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617.

Paul, D. and Johnstone, I. (2012). Augmented sparse principal component analysis for high dimensional data. *Arxiv preprint arXiv:1202.1242*.

Shen, D., Shen, H., and Marron, J. (2012). Consistency of sparse pca in high dimension, low sample size contexts. *to appear Journal of Multivariate Analysis*.

Shen, H. and Huang, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034.

Tropp, J. A. (2010). User-friendly tail bounds for sums of random matrices. *arXiv preprint arXiv:1004.4389*.

van de Geer, S. and Lederer, J. (2011). The bernstein–orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, pages 1–26.
Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027.

Vu, V. and Lei, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. Journal of Machine Learning Research (AISTATS Track).

Wang, Z., Han, F., and Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. Journal of Machine Learning Research (AISTATS Track).

Wegkamp, M. and Zhao, Y. (2013). Analysis of elliptical copula correlation factor model with kendall’s tau. Personal Communication.

Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 10:515–534.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional non-paranormal graphical models. The Annals of Statistics, 40(5):2541–2571.

Zhang, Y. and El Ghaoui, L. (2011). Large-scale sparse principal component analysis with application to text data. Proceedings of the Twenty-fourth Annual Conference on Neural Information Processing Systems.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15:265–286.