Statistica Sinica Preprint No: SS-12-052R2	
Title	On estimation of mean squared errors of benchmarked empirical Bayes estimators
Manuscript ID	SS-12-052R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2012.052
Complete List of Authors	Lingzhou Xue and Hui Zou
Corresponding Author	Hui Zou
E-mail	zouxx019@umn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Rank-based Tapering Estimation of Bandable Correlation Matrices

Lingzhou Xue and Hui Zou*

Princeton University and University of Minnesota

October 20, 2012

Abstract

The nonparanormal model assumes that variables follow a multivariate normal distribution after a set of unknown monotone increasing transformations. It is a flexible generalization of the normal model but retains the nice interpretability of the latter. In this paper we propose a rank-based tapering estimator for estimating the correlation matrix in the nonparanormal model in which the variables have a natural order. The rank-based tapering estimator does not require knowing or estimating the monotone transformation functions. We establish the rates of convergence of the rank-based tapering under Frobenious and matrix operator norms, where the dimension is allowed to grow at a nearly exponential rate as the sample size. Monte Carlo simulation is used to demonstrate the finite performance of the rank-based tapering estimator. A real data example is used to illustrate the nonparanormal model and the efficacy of the proposed rank-based tapering estimator.

Keywords: Banding, Correlation matrix, Gaussian copula, Tapering, Nonparanormal model, Variable transformation.

*Address for correspondence: Hui Zou, School of Statistics, University of Minnesota, Email: zouxx019@umn.edu. The authors thank the AE and referees for their helpful comments and suggestions. This work was completed when Lingzhou Xue was a PhD candidate at University of Minnesota. This work is supported in part by NSF grant DMS-0846068 and a grant from Office of Naval Research.

1 Introduction

Estimating large covariance matrices has been a hot research topic in recent years. High-dimensional data frequently appear in many fields, but the usual sample covariance matrix is a very poor estimator of the covariance matrix in the high-dimensional setting (Johnstone 2001). Better covariance estimators can be produced by using regularization methods, including banding (Wu & Pourahmadi 2003, Bickel & Levina 2008a), tapering (Furrer & Bengtsson 2007, Cai, Zhang & Zhou 2010, Cai & Zhou 2012) and thresholding (Bickel & Levina 2008b, El Karoui 2008, Rothman, Levina & Zhu 2010). Thresholding achieves a good variance-bias tradeoff by truncating small entries of the sample covariance matrix to zero. Thresholding is permutation invariant. On the other hand, banding and tapering utilize the underlying bandable structure of the population covariance matrix. It has shown that banding/tapering performs better than thresholding when there is a natural order among variables and two variables become near independent as they far apart in that order (Bickel & Levina 2008a, Cai et al. 2010, Cai & Zhou 2012).

The minimax results in Cai et al. (2010) and Cai & Zhou (2012) greatly deepen our understanding of the tapering estimator of large bandable covariance matrices. The minimax lower bounds were established for collections of multivariate normal distributions, which suggests that the covariance matrix estimation problem is not any easier on normal data. Technically, the rates of convergence of banding or tapering can be established under a weaker sub-Gaussian distribution assumption (Bickel & Levina 2008a,b, Cai, Zhang & Zhou 2010, Cai & Zhou 2012). Normality is important to the model interpretation, because a zero entry implies the marginal independence of a pair of variables. This nice interpretation does not always hold without the normality assumption. In practice, the observed data are often skewed or have heavy tails, especially in the high-dimensional setting. Neither normal nor sub-Gaussian distributions can be used to model such data. It is of great importance and interest to relax the normality assumption while keeping its nice interpretability.

To this end, we consider the nonparanormal model which basically uses the classical transformation strategy to handle non-normal data.

The nonparanormal model: $(X_1, \ldots, X_p)$ *follows a nonparanormal distribu-*

tion if there exists a vector of unknown univariate monotone increasing transformations, denoted by $\boldsymbol{f} = (f_1, \cdots, f_p)$*, such that the transformed random vector follows a multivariate normal distribution with mean 0 and covariance* $\boldsymbol{\Sigma}$*:*

$$(f_1(X_1), \cdots, f_p(X_p)) \sim N_p(0, \boldsymbol{\Sigma}), \tag{1}$$

where without loss of generality we let the diagonals of $\boldsymbol{\Sigma}$ all equal 1 because the transformation functions f_js can be arbitrarily scaled.

Note that $\boldsymbol{\Sigma}$ can be also regarded as the correlation matrix because the diagonals of $\boldsymbol{\Sigma}$ all equal 1. This view is very useful when developing a good estimator of $\boldsymbol{\Sigma}$ in this paper. By definition of the nonparanormal model, one can immediately see that $\boldsymbol{\Sigma}_{ij} = 0$ if and only if X_i and X_j are marginally independent. Moreover, each continuous variable can be transformed to a standard normal variable via a monotone increasing transformation. Thus, the nonparanormal model actually assumes that after these individual transformations the marginally normal variables have a joint normal distribution. This is the parametric part of model (1). The nonparanormal model is in fact a semiparametric Gaussian copula model that has generated a lot of interests in statistics, econometrics and finance (Klaassen & Wellner 1997, Song 2000, Tsukahara 2005, Chen & Fan 2006, Chen et al. 2006, Song et al. 2009). In the context of nonparametric graphical modeling, Liu, Lafferty & Wasserman (2009) used model (1) and coined the new name "nonparanormal model". We follow their terminology in this paper.

When the nonparanormal model is applied to variables with a natural order, the existing results on large covariance matrix estimation suggest us to use banding or tapering to estimate $\boldsymbol{\Sigma}$. However, an obvious difficulty is that the nonparanormal model has p many unknown nonparametric transformation functions. It appears to be inevitable that one must estimate these p transformation functions in the process of estimating $\boldsymbol{\Sigma}$. In this paper we propose a rank-based tapering estimator of $\boldsymbol{\Sigma}$ that does not require estimating these unknown transformation functions at all. Our estimator is constructed in two steps. We first construct a nonparametric rank-based sample estimate of $\boldsymbol{\Sigma}$. The rank-based tapering estimator is then obtained by applying tapering to the rank-based sample estimate of $\boldsymbol{\Sigma}$. We establish the rates of convergence of the rank-based tapering estimator under both the

Frobenius and matrix operator norms. It is shown that the rank-based tapering estimator is consistent even when the dimension is nearly exponentially large relative to the sample size.

The rest of the paper is organized as follows. Section 2 contains the methodological details of the rank-based tapering estimator. In Section 3 we present the main theoretical results. In Section 4 we use Monte Carlo simulation to demonstrate the good finite sample performance of the rank-based tapering estimator. Rock spectrum data are used to illustrate the nonparanormal model and the efficacy of the rank-based tapering estimator. Technical proofs are presented in the appendix.

2 Methodology

Throughout the rest of the paper, we assume that we have n identically independently distributed (*i.i.d.*) observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the nonparanormal model (1). Moreover, these variables follow a natural order such that banding/tapering estimation is meaningful.

We put the observed data in a $n \times p$ matrix $\boldsymbol{X}$. We can define $\boldsymbol{Z}_{ij} = f_j(\boldsymbol{X}_{ij})$, $1 \leq i \leq n; 1 \leq j \leq p$. We call $\boldsymbol{Z}$ the "oracle" data, because we would use them to estimate $\boldsymbol{\Sigma}$ if we knew the transformation functions. We begin with a key observation that $\sigma_{ij} = corr(f_i(X_i), f_j(X_j))$ for any (i, j) pair. We propose to use the rank correlation measure (Kendall 1948, Lehmann 1998) to estimate entries of $\boldsymbol{\Sigma}$. Let $(x_{1i}, x_{2i}, \ldots, x_{ni})$ be the observed values of variable X_i and denote their ranks by $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_{ni})$. We can estimate σ_{ij} by $\hat{r}^s_{ij}$ where

$$\hat{r}^s_{ij} = 2\sin(\frac{\pi}{6}\hat{r}_{ij}) \tag{2}$$

and

$$\hat{r}_{ij} = \frac{\sum_{l=1}^{n}(r_{li} - \frac{n+1}{2})(r_{lj} - \frac{n+1}{2})}{\sqrt{\sum_{l=1}^{n}(r_{li} - \frac{n+1}{2})^2 \cdot \sum_{l=1}^{n}(r_{lj} - \frac{n+1}{2})^2}}. \tag{3}$$

Note that $\hat{r}_{ij}$ is the Spearman's rank correlation and $\hat{r}^s_{ij}$ is called the adjusted Spearman's rank correlation (Kendall 1948).

It is critically important to observe that $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_{ni})$ are the ranks of the "oracle" data. Therefore, the observed data and the "oracle" data are treated as the same in the framework of rank-based estimation. The nonparanormal model implies that $f_i(X_i), f_j(X_j)$

follow a bivariate normal distribution with correlation parameter σ_{ij}. Then a classical result due to Kendall (1948) shows that

$$\lim_{n\to+\infty} \mathrm{E}(\hat{r}_{ij}) = \frac{6}{\pi} \arcsin(\frac{1}{2}\sigma_{ij}), \tag{4}$$

which indicates that the adjusted Spearman's rank correlation $\hat{r}^s_{ij}$ is an asymptotically unbiased estimator of σ_{ij}. Based on the above discussion we define the rank-based sample estimate of $\mathbf{\Sigma}$ as follows

$$\hat{\boldsymbol{R}}^s = (\hat{r}^s_{ij})_{1\le i,j\le p}. \tag{5}$$

When the dimension is large, $\hat{\boldsymbol{R}}^s$ performs poorly and we need to further consider a regularized version of the rank-based sample estimate. Banding or tapering is a very useful regularization method when the variables have a natural order and off-diagonal entries of the target covariance matrix decays to zero as they move away from the diagonal. To provide a unified treatment of banding and tapering, we consider the generalized tapering estimator defined as

$$\hat{\boldsymbol{R}}^s_{gt} = (\hat{r}^s_{ij} w_{ij})_{1\le i,j\le p},$$

where the generic tapering weights $(w_{ij})_{1\le i,j\le p}$ satisfy the following properties:

(i) $w_{ij} = 1$ for $|i-j| \le k_h = \lfloor k/2 \rfloor$,

(ii) $w_{ij} = 0$ for $|i-j| > k$.

(iii) $0 \le w_{ij} \le 1$ for $k_h < |i-j| \le k$.

Both banding and tapering weights satisfy conditions (i)–(iii). Banding weights (Bickel & Levina 2008a) $w_{ij} = 1$ for $k_h < |i-j| \le k$, while tapering weights (Cai, Zhang & Zhou 2010, Cai & Zhou 2012) equal $w^{czz}_{ij} = (k - |i-j|)/k_h$ for $k_h < |i-j| \le k$. The rank-based banding estimator is $\hat{\boldsymbol{R}}^s_b = (\hat{r}^s_{ij} I_{\{|i-j|\le k\}})_{1\le i,j\le p}$ and the rank-based tapering estimator is $\hat{\boldsymbol{R}}^s_t = (\hat{r}^s_{ij} w^{czz}_{ij})_{1\le i,j\le p}$. We consider the generalized tapering estimator because in some theoretical analysis the exact form of w_{ij} for $k_h < |i-j| \le k$ does not matter.

Before ending this section we would like to mention another "plug-in" tapering estimator of $\mathbf{\Sigma}$. One could first estimate these transformation functions and then obtain the estimated

"oracle" data as:

$$\widehat{\boldsymbol{Z}}_{ij} = \hat{f}_j(\boldsymbol{X}_{ij}), \quad 1 \le i \le n; 1 \le j \le p$$

where $\hat{f}_j$ is a good estimator of f_j. The final estimator is obtained by applying tapering to the sample covariance matrix of the estimated "oracle" data. Obviously, the "plug-in" tapering estimator requires more computations than the rank-based tapering estimator. This "plug-in" estimation idea was used by Liu et al. (2009) for estimating the inverse correlation matrix in the nonparanormal model. We discuss the theoretical advantages of the rank-based tapering estimator in Section 3.

The nonparanormal model is also interesting in the framework of graphical modeling. If $\boldsymbol{\Sigma}^{-1}$ is sparse in the sense that many entries of $\boldsymbol{\Sigma}^{-1}$ are exactly zero, then the few nonzero entries correspond to the edges in a nonparametric graphical model. Rank-based estimation techniques have been independently proposed in Xue & Zou (2012) and Liu & Wasserman (2012) for estimating sparse inverse correlation matrices of the nonparanormal model.

3 Theoretical Properties

In this section we establish theoretical properties of the rank-based generalized tapering estimators. We begin with some necessary notation and definitions. For a matrix $\boldsymbol{A} = (a_{ij})_{1\le i,j\le p}$, its Frobenius norm is defined as $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, and its matrix ℓ_q-norm is defined as the operator norm induced by the vector ℓ_q-norm,

$$\|\boldsymbol{A}\|_{\ell_v} = \sup_{\boldsymbol{u}} \frac{\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_v}}{\|\boldsymbol{u}\|_{\ell_v}}.$$

The commonly used cases are $v = 1, 2, \infty$. The ℓ_1 norm is equal to $\max_i \sum_j^p |a_{ij}|$ while the ℓ_∞ norm equals $\max_j \sum_i^p |a_{ij}|$. For a symmetric matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\ell_2}$ is the largest absolute value of its eigenvalues and $\|\boldsymbol{A}\|_{\ell_2} \le \|\boldsymbol{A}\|_{\ell_1} = \|\boldsymbol{A}\|_{\ell_\infty}$. For presentation convenience, we use c and C to denote generic constants in lower and upper bounds, respectively.

For the theoretical analysis we assume that $\boldsymbol{\Sigma}$ is in a parameter space of bandable covariance matrices. Specifically, we consider the following two parameter spaces

$$\mathcal{H}_\alpha = \{\text{data follow model (1) and } \boldsymbol{\Sigma} \text{ satisfies } |\sigma_{ij}| \le \tau_1 |i-j|^{-(\alpha+1)} \text{ for } i \ne j\}.$$

$$\mathcal{F}_\alpha = \{\text{data follow model (1) and } \mathbf{\Sigma} \text{ satisfies } \max_j \sum_{i:\ |i-j|>k} |\sigma_{ij}| \le \tau_0 k^{-\alpha},\ \text{for all } k\},$$

In both spaces α specifies the rate of decay as σ_{ij} moves away from the diagonal. We assume $p \ge n$ and $\log p \le n^\kappa$ for some constant $\kappa \in (0,1)$. Note that $\log p/n \to 0$ is necessary for establishing consistency of any estimator of $\mathbf{\Sigma}$ (Cai et al. 2010, Cai & Zhou 2012).

These two parameter spaces are similar to those considered in previous work (Bickel & Levina 2008a, Cai et al. 2010, Cai & Zhou 2012), but there is also a fundamental difference. In this work we assume the data follow a nonparanormal distribution, while the previous papers assume sub-Gaussian data.

In the following theorem we establish the rate of convergence of the rank-based (generalized) tapering estimator $\hat{\boldsymbol{R}}^s_{gt}$ under the Frobenius norm.

Theorem 1. *For the rank-based (generalized) tapering estimator $\hat{\boldsymbol{R}}^s_{gt}$ we have*

$$\sup_{\mathcal{H}_\alpha} \mathrm{E}\frac{1}{p}\big\|\hat{\boldsymbol{R}}^s_{gt} - \mathbf{\Sigma}\big\|_F^2 \le C\frac{k}{n} + Ck^{-2\alpha-1}.$$

Picking $k = n^{\frac{1}{2\alpha+2}}$, then the Frobenius risk bound becomes

$$\sup_{\mathcal{H}_\alpha} \mathrm{E}\frac{1}{p}\big\|\hat{\boldsymbol{R}}^s_{gt} - \mathbf{\Sigma}\big\|_F^2 \le Cn^{-\frac{2\alpha+1}{2(\alpha+1)}}. \tag{6}$$

Theorem 1 also suggests that the rank-based tapering estimator attains the minimax rate of convergence under the Frobenius norm. To see this, we cite a minimax lower bound from Cai et al. (2010) who constructed a special collection of multivariate distributions, denoted by $\mathcal{G}_2$, and showed that $\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathcal{G}_2} \mathrm{E}\frac{1}{p}\big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|_F^2 \ge cn^{-\frac{2\alpha+1}{2\alpha+2}}$. To save space we do not write down $\mathcal{G}_2$ here, please see section 4.2 of Cai et al. (2010) for the detail. By definition $\mathcal{G}_2$ is a subspace of $\mathcal{H}_\alpha$ and hence

$$\inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathcal{H}_\alpha} \mathrm{E}\frac{1}{p}\big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|_F^2 \ge \inf_{\hat{\mathbf{\Sigma}}} \sup_{\mathcal{G}_2} \mathrm{E}\frac{1}{p}\big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\big\|_F^2 \ge cn^{-\frac{2\alpha+1}{2\alpha+2}}. \tag{7}$$

Comparing (7) and (6) we see that the rank-based tapering estimator attains the minimax rate of convergence under the Frobenius norm.

In the next theorem we establish the rates of convergence of the rank-based (generalized) tapering estimator $\hat{\boldsymbol{R}}^s_{gt}$ under the matrix operator norms.

Theorem 2. *For the rank-based (generalized) tapering estimator* $\hat{\boldsymbol{R}}^s_{gt}$ *we have*

$$\sup_{\mathcal{F}_\alpha} \mathrm{E}\big\|\hat{\boldsymbol{R}}^s_{gt} - \boldsymbol{\Sigma}\big\|^2_{\ell_a} \le C\frac{k^2 \log p}{n} + Ck^{-2\alpha},$$

where $a = 1, 2$. *Moreover, if let* $k = (\frac{n}{\log p})^{\frac{1}{2\alpha+2}}$, *then the* ℓ_a *risk bound* ($a = 1, 2$) *becomes*

$$\sup_{\mathcal{F}_\alpha} \mathrm{E}\big\|\hat{\boldsymbol{R}}^s_{gt} - \boldsymbol{\Sigma}\big\|^2_{\ell_a} \le C\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}. \tag{8}$$

Cai et al. (2010) and Cai & Zhou (2012) have established the following lower bound results

$$\inf_{\hat{\boldsymbol{\Sigma}}} \sup_{\mathcal{G}^{**}} \mathrm{E}\big\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big\|^2_{\ell_2} \ge cn^{-2\alpha/(2\alpha+1)} + c\frac{\log(p)}{n}$$

$$\inf_{\hat{\boldsymbol{\Sigma}}} \sup_{\mathcal{G}^{*}} \mathrm{E}\big\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big\|^2_{\ell_1} \ge c\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + cn^{-\frac{\alpha}{\alpha+1}}$$

where $\mathcal{G}^{**}$ and $\mathcal{G}^{*}$ are two carefully designed collections of multivariate normal distributions. To save space we do not write down $\mathcal{G}^{**}$ and $\mathcal{G}^{*}$ here, please see section 3.2 of Cai et al. (2010) and section 2.2 of Cai & Zhou (2012) for more details.

By definition, both $\mathcal{G}^{**}$ and $\mathcal{G}^{*}$ are subspaces of $\mathcal{F}_\alpha$. As a result, we have

$$\inf_{\hat{\boldsymbol{\Sigma}}} \sup_{\mathcal{F}_\alpha} \mathrm{E}\big\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big\|^2_{\ell_2} \ge cn^{-2\alpha/(2\alpha+1)} + c\frac{\log(p)}{n},$$

$$\inf_{\hat{\boldsymbol{\Sigma}}} \sup_{\mathcal{F}_\alpha} \mathrm{E}\big\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\big\|^2_{\ell_1} \ge c\left(\frac{\log p}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + cn^{-\frac{\alpha}{\alpha+1}}.$$

So Theorem 2 does not tell us whether the rank-based (generalized) tapering estimator is minimax rate optmial under the ℓ_1, ℓ_2 norms. Because we are dealing with the rank correlations, some key inequalities used in establishing the upper bound for the ℓ_1, ℓ_2 risk are no longer applicable. For example, the concentration bound for sub-Gaussian random variables (see page 2142 of Cai et al. (2010) and Saulis & Statulevičius (1991)). The proof of Theorem 2 uses a generalization of McDiarmid's inequality by Kutin (2002) and Kutin & Niyogi (2002).

With or without the minimax optimality, Theorems 1 and 2 show that the rank-based tapering estimator is uniformly consistent over a large parameter space, as long as the logarithm of the dimension grows slower than the sample size. Let us compare the rank-based tapering estimator and the "plug-in" tapering estimator. Liu et al. (2009) used a

nonparametric density estimation scheme to estimate the transformation functions in the nonparanormal model and used the same "plug-in" idea to estimate $\mathbf{\Sigma}$ by applying the graphical lasso (Friedman, Hastie & Tibshirani 2008) to the estimated "oracle" data. Their theory is based on a concentration inequality which is proven under the assumption that p is a polynomial order of n. So their theory may suggest the the "plug-in" tapering estimator is consistent under polynomial dimensions but it is unclear whether the "plug-in" tapering estimator can still be consistent under nearly exponentially large dimensions.

4 Numerical Properties

In this section we both simulated data and real data to examine the finite-sample performance of the proposed rank-based tapering estimator.

4.1 Monte Carlo Simulation

The main purpose of the simulation study is to show that the proposed rank-based tapering estimator works as well as the oracle tapering estimator. For the sake of completeness we also include the "plug-in" estimator for comparison.

We generated n independent p-dimension data points from the nonparanormal model (1) with $n = 200$ and $p = 200,\ 500$ & 1000. Four different $\mathbf{\Sigma}$ were considered:

1. $\sigma_{ij} = I_{\{i=j\}} + \rho|i-j|^{-(\alpha+1)} I_{\{i\neq j\}}$ with $\rho = 0.6$, and $\alpha = 0.1, 0.3, 0.5$.

2. $\sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.3$ and 0.7.

3. $\sigma_{ij} = I_{\{i=j\}} + \rho I_{\{|i-j|=1\}}$ for $\rho = 0.3$ and 0.5.

4. $\sigma_{ij} = (1 - \frac{|i-j|}{m})_+$ for $m = 0.1p$, $0.2p$ and $0.3p$.

Model 1 has been used in Cai et al. (2010). Models 2 and 3 have been used in Bickel & Levina (2008a,b) and Rothman et al. (2009). Model 4 has been studied in Cai & Liu (2011). We first generated n independent data from $N_p(0, \mathbf{\Sigma})$ and then transformed the normal data using transformation functions. In the simulation study we considered two

sets of transformation functions for each $\mathbf{\Sigma}$. We applied the identity transformation to obtain the normal data. We also generated nonparanormal data by applying the following transformations

$$\boldsymbol{g} = [f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_5^{-1}, f_1^{-1}, f_2^{-1}, f_3^{-1}, f_4^{-1}, f_5^{-1}, \ldots],$$

where $f_1(x) = x$, $f_2(x) = \log(x)$, $f_3(x) = x^{\frac{1}{3}}$ and $f_4(x) = \log(\frac{x}{1-x})$.

The estimators considered in the study are the direct banding/tapering estimator, the proposed rank-based banding/tapering estimator and the "plug-in" estimator. See Table 1. To construct the "plug-in" estimator, we first estimated the "oracle" data $\hat{\boldsymbol{z}}_i = \widehat{\boldsymbol{f}}(\boldsymbol{x}_i)$ by applying the estimated transformation vector $\widehat{\boldsymbol{f}} = (\hat{f}_1, \cdots, \hat{f}_p) = (\Phi^{-1} \circ \hat{F}_1, \cdots, \Phi^{-1} \circ \hat{F}_p)$ with $\hat{F}_j$ being a Winsorized estimator of the CDF of X_j (Liu et al. 2009), and then performed the banding/tapering procedure over the estimated "oracle" data. By our simulation design, no matter which transformation function is used to generate the nonparanormal data, the "oracle" data are always the normal data on which the transformation is applied. Therefore, although the direct banding/tapering estimator is obviously wrong for the nonparanormal data, their results on the normal data are actually the results of the ideal banding/tapering estimator for both normal and nonparanormal data cases, which can be used as the benchmark for comparison. We only report the results of direct banding/tapering for the normal data. Each estimator is tuned by 5-fold cross-validation. The estimation accuracy is measured by the average ℓ_1-norm over 100 independent replications.

The simulation results are summarized in Tables 2–5. We can draw several conclusions. First, the rank-based banding/tapering estimators work very similarly to the ideal banding/tapering estimator, whose results correspond to those by the direct banding/tapering estimator on normal data. The rank-based estimator is only slightly worse than the ideal estimator, which is expected because some information is lost in the process of converting the original data into their ranks. The rank-based estimators outperform the "plug-in" estimators. We also compared these estimators using Frobenius norm and ℓ_2 norm. The conclusions stay the same. For the sake of space we do not present the simulation results under Frobenius norm and ℓ_2 norm here.

Table 1: List of all estimators in our simulation study.

Notation	Meaning
$\hat{\mathbf{\Sigma}}_b^d$	the direct banding estimator
$\hat{\mathbf{\Sigma}}_t^d$	the direct tapering estimator
$\hat{\boldsymbol{R}}_b^s$	the rank-based banding estimator
$\hat{\boldsymbol{R}}_t^s$	the rank-based tapering estimator
$\hat{\mathbf{\Sigma}}_b^p$	the "plug-in" banding estimator
$\hat{\mathbf{\Sigma}}_t^p$	the "plug-in" tapering estimator

4.2 Applications to the Rock Spectrum Data

We use the rock sonar spectrum data (Gorman & Sejnowski 1988) to illustrate the nonparanormal model and the efficacy of the proposed rank-based banding/tapering estimator. This dataset consists of 97 sonar spectra bounced off from roughly cylindrical rocks under similar conditions, and each spectrum has 60 frequency band energy measurements in the range 0.0 to 1.0. First, we conducted normality tests on these 60 spectra signals to check whether there is a serious violation of normality. The testing results are reported in Table 6. More than 80% signals are unable to pass any of four normality tests, and under Bonferroni correction there are still over 50% genes that fail to pass the four normality tests. The normality test strongly suggests that the normal assumption does not hold for the rock spectrum data. Hence we considered the more robust nonparanormal model for this dataset.

For each spectrum, these 60 spectra signals were obtained from an increasing order of 60 aspect angles spanning 180 degrees. Thus, there is a natural order among signals for each spectra. This physical nature of the data motivates us to estimate its bandable correlation matrix structure. Moreover, the heatmap of the Spearman's correlation matrix is shown in Figure 1, The heatmap shows a general decaying pattern, which suggests that it is quite reasonable to assume the correlation matrix of the nonparanormal model is bandable. After the visual inspection, we computed the rank-based banding/tapering estimators. The rank-based banding estimator selected $\hat{k} = 27$ sub-diagonals by cross-validation, while the rank-based tapering estimator selected $\hat{k} = 38$ sub-diagonals. The heatmaps of the absolute values

Table 2: Simulation results of model 1. Estimation accuracy is measured by the ℓ_1-norm averaged over 100 replications. The standard errors are shown in the bracket. In this simulation study, the direct banding/tapering estimator on the normal data corresponds to an ideal banding/tapering estimator on both normal and nonparanormal data.

α	0.1			0.3			0.5		
p	200	500	1000	200	500	1000	200	500	1000
					normal data				
$\hat{\boldsymbol{\Sigma}}_b^d$	3.50	4.27	4.78	2.35	2.67	2.86	1.68	1.83	1.92
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
$\hat{\boldsymbol{R}}_b^s$	3.55	4.33	4.82	2.39	2.71	2.90	1.71	1.86	1.95
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
$\hat{\boldsymbol{\Sigma}}_b^p$	3.57	4.34	4.88	2.43	2.75	2.95	1.76	1.92	2.01
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
$\hat{\boldsymbol{\Sigma}}_t^d$	3.40	4.19	4.71	2.28	2.60	2.80	1.64	1.81	1.92
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
$\hat{\boldsymbol{R}}_t^s$	3.46	4.25	4.76	2.32	2.65	2.85	1.68	1.84	1.94
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
$\hat{\boldsymbol{\Sigma}}_t^p$	3.52	4.30	4.84	2.39	2.72	2.93	1.75	1.91	2.02
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
					nonparanormal data				
$\hat{\boldsymbol{R}}_b^s$	3.55	4.33	4.82	2.39	2.71	2.90	1.71	1.86	1.95
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
$\hat{\boldsymbol{\Sigma}}_b^p$	3.57	4.34	4.88	2.43	2.75	2.95	1.76	1.92	2.01
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
$\hat{\boldsymbol{R}}_t^s$	3.46	4.25	4.76	2.32	2.65	2.85	1.68	1.84	1.94
	(0.01)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)
$\hat{\boldsymbol{\Sigma}}_t^p$	3.52	4.30	4.84	2.39	2.72	2.93	1.75	1.91	2.02
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)

Table 3: Simulation results of model 2. Estimation accuracy is measured by the ℓ_1-norm averaged over 100 replications. The standard errors are shown in the bracket. In this simulation study, the direct banding/tapering estimator on the normal data corresponds to an ideal banding/tapering estimator on both normal and nonparanormal data.

ρ	0.3			0.7		
p	200	500	1000	200	500	1000
normal data						
$\hat{\boldsymbol{\Sigma}}_b^d$	0.57	0.61	0.64	1.74	1.85	1.94
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
$\hat{\boldsymbol{R}}_b^s$	0.59	0.63	0.67	1.80	1.91	2.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{\Sigma}}_b^p$	0.62	0.66	0.69	1.89	2.02	2.11
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{\Sigma}}_t^d$	0.59	0.62	0.66	1.67	1.80	1.90
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{R}}_t^s$	0.60	0.62	0.63	1.74	1.87	1.96
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{\Sigma}}_t^p$	0.63	0.67	0.70	1.87	2.01	2.10
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
nonparanormal data						
$\hat{\boldsymbol{R}}_b^s$	0.59	0.63	0.67	1.80	1.91	2.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{\Sigma}}_b^p$	0.62	0.66	0.69	1.89	2.02	2.11
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{R}}_t^s$	0.60	0.62	0.63	1.74	1.87	1.96
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)
$\hat{\boldsymbol{\Sigma}}_t^p$	0.63	0.67	0.70	1.87	2.01	2.10
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)

Table 4: Simulation results of model 3. Estimation accuracy is measured by the ℓ_1-norm averaged over 100 replications. The standard errors are shown in the bracket. In this simulation study, the direct banding/tapering estimator on the normal data corresponds to an ideal banding/tapering estimator on both normal and nonparanormal data.

ρ		0.3			0.5	
p	200	500	1000	200	500	1000
			normal data			
$\hat{\boldsymbol{\Sigma}}_b^d$	0.29	0.31	0.34	0.24	0.27	0.28
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{R}}_b^s$	0.31	0.33	0.35	0.26	0.29	0.30
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{\Sigma}}_b^p$	0.35	0.37	0.39	0.32	0.34	0.35
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{\Sigma}}_t^d$	0.29	0.31	0.34	0.24	0.27	0.28
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{R}}_t^s$	0.31	0.33	0.35	0.26	0.29	0.30
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{\Sigma}}_t^p$	0.35	0.37	0.39	0.32	0.34	0.35
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
			nonparanormal data			
$\hat{\boldsymbol{R}}_b^s$	0.31	0.33	0.35	0.26	0.29	0.30
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{\Sigma}}_b^p$	0.35	0.37	0.39	0.32	0.34	0.35
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{R}}_t^s$	0.31	0.33	0.35	0.26	0.29	0.30
	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)
$\hat{\boldsymbol{\Sigma}}_t^p$	0.35	0.37	0.39	0.32	0.34	0.35
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)

Table 5: Simulation results of model 4. Estimation accuracy is measured by the ℓ_1-norm averaged over 100 replications. The standard errors are shown in the bracket. In this simulation study, the direct banding/tapering estimator on the normal data corresponds to an ideal banding/tapering estimator on both normal and nonparanormal data.

m	0.1p			0.2p			0.3p		
p	200	500	1000	200	500	1000	200	500	1000
					normal data				
$\hat{\boldsymbol{\Sigma}}_b^d$	1.55	3.99	8.06	2.87	7.52	15.58	5.20	13.70	28.33
	(0.02)	(0.08)	(0.19)	(0.05)	(0.15)	(0.56)	(0.11)	(0.41)	(1.14)
$\hat{\boldsymbol{R}}_b^s$	1.67	4.22	8.82	3.08	8.13	16.64	5.55	14.57	30.59
	(0.03)	(0.09)	(0.21)	(0.05)	(0.17)	(0.55)	(0.10)	(0.41)	(1.26)
$\hat{\boldsymbol{\Sigma}}_b^p$	1.88	4.77	9.86	3.48	9.26	18.47	6.47	16.44	33.92
	(0.03)	(0.08)	(0.22)	(0.06)	(0.29)	(0.57)	(0.11)	(0.44)	(1.23)
$\hat{\boldsymbol{\Sigma}}_t^d$	1.63	4.19	8.56	3.07	8.02	16.50	5.25	13.70	27.97
	(0.03)	(0.09)	(0.22)	(0.06)	(0.21)	(0.61)	(0.09)	(0.35)	(1.11)
$\hat{\boldsymbol{R}}_t^s$	1.73	4.40	9.25	3.20	8.48	17.41	5.64	14.62	30.65
	(0.03)	(0.10)	(0.22)	(0.05)	(0.20)	(0.64)	(0.10)	(0.37)	(1.29)
$\hat{\boldsymbol{\Sigma}}_t^p$	2.06	5.21	10.79	3.80	10.12	20.20	6.85	17.47	36.09
	(0.03)	(0.09)	(0.24)	(0.06)	(0.27)	(0.63)	(0.13)	(0.41)	(1.36)
					nonparanormal data				
$\hat{\boldsymbol{R}}_b^s$	1.67	4.22	8.82	3.08	8.13	16.64	5.55	14.57	30.59
	(0.03)	(0.09)	(0.21)	(0.05)	(0.17)	(0.55)	(0.10)	(0.41)	(1.26)
$\hat{\boldsymbol{\Sigma}}_b^p$	1.88	4.77	9.86	3.48	9.26	18.47	6.47	16.44	33.92
	(0.03)	(0.08)	(0.22)	(0.06)	(0.29)	(0.57)	(0.11)	(0.44)	(1.23)
$\hat{\boldsymbol{R}}_t^s$	1.73	4.40	9.25	3.20	8.48	17.41	5.64	14.62	30.65
	(0.03)	(0.10)	(0.22)	(0.05)	(0.20)	(0.64)	(0.10)	(0.37)	(1.29)
$\hat{\boldsymbol{\Sigma}}_t^p$	2.06	5.21	10.79	3.80	10.12	20.20	6.85	17.47	36.09
	(0.03)	(0.09)	(0.24)	(0.06)	(0.27)	(0.63)	(0.13)	(0.41)	(1.36)

Table 6: Normality test results for the rock spectrum data. The counts of spectra that fail to pass each normality test are shown in the table.

critical value	Anderson–Darling	Cramér–von Mises	Lilliefors	Shapiro-Francia
0.05	55	52	48	56
0.05/60	45	41	31	43

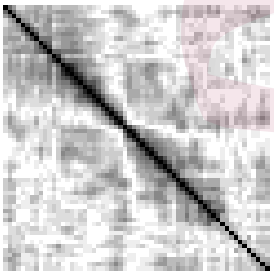

Figure 1: Heapmaps of the absolute values of the Pearson's and Spearman's correlation matrices for the rock spectrum data. White means zero correlation and black means perfect correlation (magnitude equals 1).

of the rank-based banding/tapering estimators are shown in Figure 2.

It is interesting to note that under the nonparanormal model, if the correlation matrix is exactly banded, then the direct correlation matrix of the raw data will be exactly banded too. Now that we have came to a conclusion that a nonparanormal model with an exactly banded correlation matrix is a good fit to the rock spectrum data, then we would expect the similar bandable structure should also hold for the direct correlation matrix of the raw data. We performed the direct banding/tapering procedures on the raw data and we did obtain the same bandable structure: the direct banding chose $\hat{k} = 27$ sub-diagonals and the direct tapering selected $\hat{k} = 38$ sub-diagonals. We show the heatmaps of the absolute values of the direct banding/tapering estimators in Figure 3, providing another support to the fitted nonparanormal model.

(A): rank-based banding estimation.

(B): rank-based tapering estimation.

Figure 2: Heapmaps of the absolute values of the rank-based banding/tapering estimator for the rock spectrum data.

(A): direct banding estimation on the raw data.

(B): direct tapering estimation on the raw data.

Figure 3: Heapmaps of the absolute values of the direct banding/tapering estimator for the rock spectrum data.

5 Discussion

In this paper we have introduced a rank-based generalized tapering estimator for estimating a high-dimensional correlation matrix in the nonparanormal model. The theoretical and numerical examples have provided strong support for this estimation method. Tapering estimation requires a natural order among the variables. If there is no such order information, it is better to use a permutation invariance estimator such as thresholding. In Xue & Zou (2011) we have proved the adaptive minimax optimality of a rank-based thresholding estimator for estimating sparse correlation matrices of nonparanormal models.

Appendix: technical proofs

We first present a useful technical lemma concentration bounds concerning the accuracy of the rank-based sample estimator. Its proof is given in Xue & Zou (2012).

Lemma 1. *Fix any* $0 < \varepsilon < 1$ *and let* $n \geq \frac{12\pi}{\varepsilon}$. *We have*

$$\Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \varepsilon) \leq 2\exp(-c_0 n\varepsilon^2),$$

$$\Pr(\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}\|_{max} > \varepsilon) \leq p^2\exp(-c_0 n\varepsilon^2),$$

where $\|\hat{\boldsymbol{R}}^s - \boldsymbol{\Sigma}\|_{max} = \max_{(i,j)}|\hat{r}^s_{ij} - \sigma_{ij}|$ *is the max norm, and* c_0 *is some absolute constant.*

Proof of Theorem 1. Introduce $\boldsymbol{\Gamma}_b = (\sigma_{ij}I_{\{|i-j|\leq k\}})_{1\leq i,j\leq p}$ and $\boldsymbol{\Gamma}_{gt} = (\sigma_{ij}w_{ij})_{1\leq i,j\leq p}$ and we have

$$\mathrm{E}\big\|\hat{\boldsymbol{R}}^s_{gt} - \boldsymbol{\Sigma}\big\|^2_F \leq \mathrm{E}\big\|\hat{\boldsymbol{R}}^s_{gt} - \boldsymbol{\Gamma}_{gt}\big\|^2_F + \big\|\boldsymbol{\Gamma}_{gt} - \boldsymbol{\Gamma}_b\big\|^2_F + \big\|\boldsymbol{\Gamma}_b - \boldsymbol{\Sigma}\big\|^2_F.$$

By the assumption $|\sigma_{ij}| \leq \tau_1|i-j|^{-(\alpha+1)}$ for $i \neq j$ and $0 \leq w_{ij} \leq 1$, it immediately yields that $\big\|\boldsymbol{\Gamma}_{gt} - \boldsymbol{\Gamma}_b\big\|^2_F \leq Cpk^{-2\alpha-2}$. Next, we can also derive the upper bound for $\big\|\boldsymbol{\Gamma}_b - \boldsymbol{\Sigma}\big\|^2_F$ as follows,

$$\big\|\boldsymbol{\Gamma}_b - \boldsymbol{\Sigma}\big\|^2_F \leq 2\tau_1 \cdot \sum_{m=k+1}^{p-1} m^{-2(\alpha+1)} \leq 2\tau_1 \cdot pk^{-2\alpha-1}. \tag{9}$$

where we use the fact that

$$\sum_{m=k+1}^{p-1} m^{-2(\alpha+1)} \leq \int_k^{+\infty} t^{-2(\alpha+1)}dt \leq \frac{1}{2\alpha}k^{-2\alpha-1}.$$

On the other hand, $(\hat{r}_{ij}^s - \sigma_{ij})^2 \le \frac{2\pi^2}{9}(u_{ij} - E(u_{ij}))^2 + O(\frac{1}{n^2})$ holds by the Hoeffding decomposition of $\hat{r}_{ij}$. Note that $\mathrm{Var}(u_{ij}) = O(\frac{1}{n})$, and then it can be easily seen that

$$\mathrm{E}\frac{1}{p}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}\big\|_F^2 \le \mathrm{E}\frac{1}{p}\big\|\hat{\boldsymbol{R}}_b^s - \boldsymbol{\Gamma}_b\big\|_F^2 \le \frac{2\pi^2}{9}k \cdot (\max_{i,j}\mathrm{Var}(u_{ij}) + \frac{c}{n^2}) \le C\frac{k}{n}. \tag{10}$$

Combining (9) and (10) concludes the proof by noting that

$$\mathrm{E}\frac{1}{p}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Sigma}\big\|_F^2 \le C\frac{k}{n} + Ck^{-2\alpha-1}. \qquad \square$$

Proof of Theorem 2. We only need to prove the ℓ_1 risk bound. The ℓ_2 risk bound follows from the fact that the ℓ_2 norm is upper bounded by the ℓ_1 norm. Introduce the ideal generalized tapering estimator $\boldsymbol{\Gamma}_{gt} = (\sigma_{ij}w_{ij})_{1\le i,j\le p}$. The sub-additive property of the matrix ℓ_1-norm implies that

$$\mathrm{E}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Sigma}\big\|_{\ell_1}^2 \le 2\mathrm{E}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}\big\|_{\ell_1}^2 + 2\big\|\boldsymbol{\Gamma}_{gt} - \boldsymbol{\Sigma}\big\|_{\ell_1}^2. \tag{11}$$

Note that $0 \le w_{ij} \le 1$ for any (i,j) and $w_{ij} = 1$ when $|i-j| \le k_h$, and then we have

$$\big\|\boldsymbol{\Gamma}_{gt} - \boldsymbol{\Sigma}\big\|_{\ell_1}^2 \le \big(\max_{i=1,\ldots,p}\sum_{i:\ |i-j|>k_h}|\sigma_{ij}|\big)^2 \le \tau_0^2 k_h^{-2\alpha}.$$

Now, we only need to bound $\mathrm{E}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}\big\|_{\ell_1}^2$ in the sequel. Since $0 \le w_{ij} \le 1$ for any (i,j) and $w_{ij} = 0$ for $|i-j| > k$, we can easily derive its upper bound as follows,

$$\begin{aligned}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}\big\|_{\ell_1}^2 &\le \max_{1\le i\le p}\Big(\sum_{i-k\le j\le i+k}|\hat{r}_{ij}^s - \sigma_{ij}|\Big)^2 \\ &\le (2k+1)\cdot\max_{1\le i\le p}\sum_{i-k\le j\le i+k}(\hat{r}_{ij}^s - \sigma_{ij})^2.\end{aligned}$$

Recall that $\hat{r}_{ij}^s = 2\sin(\frac{\pi}{6}\hat{r}_{ij})$ and $\sigma_{ij} = 2\sin(\frac{\pi}{6}E(u_{ij}))$. To further simplify the upper bound, we consider the Hoeffding decomposition of $\hat{r}_{ij}$ (Hoeffding 1948),

$$\begin{aligned}(\hat{r}_{ij}^s - \sigma_{ij})^2 &\le \frac{\pi^2}{9}(\hat{r}_{ij} - E(u_{ij}))^2 \\ &\le \frac{2\pi^2}{9}\big[(u_{ij} - E(u_{ij}))^2 + (\frac{3}{n+1}u_{ij} - \frac{3}{n^3-n}d_{ij})^2\big] \\ &= \frac{2\pi^2}{9}(u_{ij} - E(u_{ij}))^2 + O(\frac{1}{n^2})\end{aligned}$$

where $d_{ij} = \sum_{i_1\ne i_2}\mathrm{sign}(x_{i_1 i} - x_{i_2 i})\cdot\mathrm{sign}(x_{i_1 j} - x_{i_2 j})$.

Now we shall derive the concentration bound for

$$U_i^{(k)} = \sum_{i-k \le j \le i+k} (u_{ij} - E(u_{ij}))^2.$$

To obtain the desired concentration bound, we consider the generalization of McDiarmid's inequality by Kutin (2002) and Kutin & Niyogi (2002) when differences are bounded with high probability. Note that $U_i^{(k)}$ can be considered as a function of independent samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, i.e. $U_i^{(k)} = U_i^{(k)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. We replace the t-th sample $\boldsymbol{x}_t$ by another independent sample $\tilde{\boldsymbol{x}}_t$. To simplify notation, we define $\tilde{U}_i^{(k)} = U_i^{(k)}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}, \tilde{\boldsymbol{x}}_t, \boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_n)$ and $\tilde{u}_{ij} = u_{ij}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}, \tilde{\boldsymbol{x}}_t, \boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_n)$. We have that

$$|U_i^{(k)} - \tilde{U}_i^{(k)}| \le \sum_{i-k \le j \le i+k} |(u_{ij} - E(u_{ij}))^2 - (\tilde{u}_{ij} - E(\tilde{u}_{ij}))^2| \le 12 \sum_{i-k \le j \le i+k} |u_{ij} - \tilde{u}_{ij}| \le c_1 \frac{k}{n},$$

where c_1 is some absolute constant. We use simple facts that $|u_{ij}| \le 3$ and $|\tilde{u}_{ij}| \le 3$ in the first inequality and $|u_{ij} - \tilde{u}_{ij}| \le \frac{15}{n}$ (Xue & Zou 2012) in the last inequality. Moreover, under the probability event that

$$\{\max_{i,j} |u_{ij} - E(u_{ij})| \le \epsilon\} \cap \{\max_{i,j} |\tilde{u}_{ij} - E(\tilde{u}_{ij})| \le \epsilon\} \text{ with } \epsilon^2 = c_0 M \frac{\log p}{n}$$

for $M > 0$, $U_i^{(k)} - \tilde{U}_i^{(k)}$ can be upper bounded as follows

$$|U_i^{(k)} - \tilde{U}_i^{(k)}| \le \sum_{i-k \le j \le i+k} (|u_{ij} - E(u_{ij})| + |\tilde{u}_{ij} - E(u_{ij})|) \cdot |u_{ij} - \tilde{u}_{ij}| \le c_2 M \frac{k \log^{1/2} p}{n^{3/2}},$$

where c_2 is some absolute constant. Thus, the following bound immediately holds

$$\Pr(|U_i^{(k)} - \tilde{U}_i^{(k)}| > c_2 M \frac{k \log p}{n^{3/2}}) \le 2 \Pr(\max_{ij} |u_{ij} - E(u_{ij})| > \epsilon) \le 4p^{-M}$$

Using the terminology from Kutin (2002) and Kutin & Niyogi (2002), we have proved that $U_i^{(k)}$ is strongly difference-bounded by

$$(b_*, c_*, \delta_*) \equiv (c_1 \frac{k}{n}, c_2 M \frac{k \log^{1/2} p}{n^{3/2}}, 4p^{-M}).$$

Next we directly apply Theorem 2.13 in Kutin & Niyogi (2002), which is a simplified version of Theorem 1.9 in Kutin (2002), to obtain the probability bound

$$\begin{aligned} \Pr(U_i^{(k)} - E(U_i^{(k)}) > \frac{x^2}{n}) &\le \exp(\frac{-x^4/n^2}{8nc_*^2}) + \frac{nb_*\delta_*}{c_*} \\ &= \exp(\frac{-x^4}{8c_2^2 M^2 k^2 \log p}) + \frac{4c_1}{c_2 M} \cdot n^{3/2} p^{-M} \log^{-1/2} p \end{aligned}$$

Since (5.13) in Hoeffding (1948) showed that $\mathrm{Var}(u_{ij}) = O(\frac{1}{n})$ for any (i,j), we then have $E[U_i^{(k)}] = O(\frac{k}{n})$. Now applying the union bound yields that for some constant $C_0 > 0$

$$\Pr\big(\max_{1\le i\le p} U_i^{(k)} > C_0(\frac{k}{n} + \frac{x^2}{n})\big) \le p \cdot \max_{1\le i\le p} \Pr(U_i^{(k)} - E(U_i^{(k)}) > \frac{x^2}{n})$$

Now, we are ready to bound the expected ℓ_1-norm of $\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}$ by truncation of $\max_{1\le i\le p} U_i^{(k)}$,

$$\begin{aligned}
\mathrm{E}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Gamma}_{gt}\big\|_{\ell_1}^2 &\le C\frac{k^2}{n^2} + \frac{2\pi^2}{9}(2k+1)\mathrm{E}(\max_{1\le i\le p} U_i^{(k)}) \\
&\le C\frac{k^2}{n^2} + Ck \cdot \mathrm{E}[\max_{1\le i\le p} U_i^{(k)}(I_{\{\max_{1\le i\le p} U_i^{(k)} \le C_0(\frac{k}{n}+\frac{x^2}{n})\}} + I_{\{\max_{1\le i\le p} U_i^{(k)} > C_0(\frac{k}{n}+\frac{x^2}{n})\}})] \\
&\le C\frac{k^2}{n^2} + C(\frac{k^2}{n} + \frac{kx^2}{n}) + Ck^2 \cdot \Pr\big(\max_{1\le i\le p} U_i^{(k)} > C(\frac{k}{n} + \frac{x^2}{n})\big) \\
&\le C_1(\frac{k^2}{n} + \frac{kx^2}{n}) + C_2k^2p\exp(-\frac{x^4}{8c_2^2M^2k^2\log p}) + C_3k^2n^{3/2}p^{-M+1}(\log p)^{-1/2}
\end{aligned}$$

where we use the fact that $|U_i^{(k)}| = O(k)$ in the last but one inequality. Thus we choose $M > 4$, and further set $x^4 = 8c_2^2M^3 \cdot k^2\log^2 p$ to conclude that

$$\mathrm{E}\big\|\hat{\boldsymbol{R}}_{gt}^s - \boldsymbol{\Sigma}\big\|_{\ell_1}^2 \le C\frac{k^2\log p}{n} + Ck^{-2\alpha}.$$

□

References

Bickel, P. & Levina, E. (2008a), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* **36**(1), 199–227.

Bickel, P. & Levina, E. (2008b), 'Covariance regularization by thresholding', *The Annals of Statistics* **36**(6), 2577–2604.

Cai, T. & Liu, W. (2011), 'Adaptive thresholding for sparse covariance matrix estimation', *Journal of the American Statistical Association* **106**(494), 672–684.

Cai, T., Zhang, C. & Zhou, H. (2010), 'Optimal rates of convergence for covariance matrix estimation', *The Annals of Statistics* **38**(4), 2118–2144.

Cai, T. & Zhou, H. (2012), 'Minimax estimation of large covariance matrices under ℓ_1-norm (with discussion)', *Statistica Sinica* **22**(4), 1319–1378.

Chen, X. & Fan, Y. (2006), 'Estimation of copula-based semiparametric time series models', *Journal of Econometrics* **130**(2), 307–335.

Chen, X., Fan, Y. & Tsyrennikov, V. (2006), 'Efficient estimation of semiparametric multivariate copula models', *Journal of the American Statistical Association* **101**(475), 1228–1240.

El Karoui, N. (2008), 'Operator norm consistent estimation of large dimensional sparse covariance matrices', *The Annals of Statistics* **36**(6), 2717–2756.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**(3), 432–441.

Furrer, R. & Bengtsson, T. (2007), 'Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants', *Journal of Multivariate Analysis* **98**(2), 227–255.

Gorman, R. & Sejnowski, T. (1988), 'Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets', *Neural Networks* **1**, 75–89.

Hoeffding, W. (1948), 'A class of statistics with asymptotically normal distribution', *The Annals of Mathematical Statistics* **19**(3), 293–325.

Johnstone, I. (2001), 'On the distribution of the largest eigenvalue in principal components analysis', *The Annals of Statistics* **29**(2), 295–327.

Kendall, M. (1948), *Rank Correlation Methods*, Charles Griffin and Co. Ltd., London.

Klaassen, C. & Wellner, J. (1997), 'Efficient estimation in the bivariate normal copula model: normal margins are least favourable', *Bernoulli* **3**(1), 55–77.

Kutin, S. (2002), Extensions to mcdiarmid's inequality when differences are bounded with high probability, Technical report, TR-2002-04, University of Chicago.

Kutin, S. & Niyogi, P. (2002), Almost-everywhere algorithmic stability and generalization error, *in* 'Proceedings of Unvertainty in Artifical Intelligence', University of Alberta, Edmonton, Canada.

Lehmann, E. (1998), *Nonparametrics: statistical methods based on ranks*, Prentice Hall Upper Saddle River, New Jersey.

Liu, H., H. F. Y. M. L. J. & Wasserman, L. (2012), 'High dimensional semiparametric Gaussian copula graphical models', *The Annals of Statistics., to appear* .

Liu, H., Lafferty, J. & Wasserman, L. (2009), 'The nonparanormal: semiparametric estimation of high dimensional undirected graphs', *Journal of Machine Learning Research* **10**, 1–37.

Rothman, A., Levina, E. & Zhu, J. (2009), 'Generalized thresholding of large covariance matrices', *Journal of the American Statistical Association* **104**(485), 177–186.

Rothman, A., Levina, E. & Zhu, J. (2010), 'A new approach to Cholesky-based covariance regularization in high dimensions', *Biometrika* **97**(3), 539.

Saulis, L. & Statulevičius, V. (1991), *Limit Theorems for Large Deviations*, Springer, Berlin.

Song, P. (2000), 'Multivariate dispersion models generated from gaussian copula', *Scandinavian Journal of Statistics* **27**(2), 305–320.

Song, P., Li, M. & Yuan, Y. (2009), 'Joint regression analysis of correlated data using gaussian copulas', *Biometrics* **65**, 60–68.

Tsukahara, H. (2005), 'Semiparametric estimation in copula models', *Canadian Journal of Statistics* **33**(3), 357–375.

Wu, W. & Pourahmadi, M. (2003), 'Nonparametric estimation of large covariance matrices of longitudinal data', *Biometrika* **90**(4), 831–844.

Xue, L. & Zou, H. (2011), 'On Optimal Estimation of Sparse Correlation Matrices of Semiparametric Gaussian Copulas', *Manuscript, submitted* .

Xue, L. & Zou, H. (2012), 'Regularized Rank-based Estimation of High-dimensional Nonparanormal Graphical Models', *The Annals of Statistics., to appear* .