Statistics and Its Interface Volume 7 (2014) 201–209

Optimal estimation of sparse correlation matrices of semiparametric Gaussian copulas

Lingzhou Xue* and Hui Zou†

Statistical inference of semiparametric Gaussian copulas is well studied in the classical fixed dimension and large sample size setting. Nevertheless, optimal estimation of the correlation matrix of semiparametric Gaussian copula is understudied, especially when the dimension can far exceed the sample size. In this paper we derive the minimax rate of convergence under the matrix ℓ_1-norm and ℓ_2-norm for estimating large correlation matrices of semiparametric Gaussian copulas when the correlation matrices are in a weak ℓ_q ball. We further show that an explicit rank-based thresholding estimator adaptively attains minimax optimal rate of convergence simultaneously for all $0 \le q < 1$. Numerical examples are provided to demonstrate the finite sample performance of the rank-based thresholding estimator.

Keywords and phrases: Correlation matrix, Gaussian copula, Minimax optimality, Rank correlation, Thresholding, Weak ℓ_q ball.

1. INTRODUCTION

Practitioners often take variable transformation before applying the intended multivariate analysis method. For example, when doing principal component analysis the correlation matrix is preferred over the covariance matrix if variables have very different scales. Using the correlation matrix is amount to using the covariance matrix of linearly transformed variables such that after transformation the mean is zero and variable is one. From this perspective, semiparametric Gaussian copulas adopt nonparametric transformation techniques and assume normality after transformation. More specifically, we have the following definition.

> *The semiparametric Gaussian copula model:* $(X_1, \ldots, X_p)'$ obeys a semiparametric Gaussian copula model with the correlation matrix $\mathbf{\Sigma}$, if there exists a vector of unknown univariate monotone increasing transformations denoted by $(f_1, \ldots, f_p)$ such that the transformed random vector follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\mathbf{\Sigma}$:
>
> $$(1)\qquad (f_1(X_1), \ldots, f_p(X_p)) \sim N_p(\mathbf{0}, \mathbf{\Sigma})',$$
>
> where $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ and $\sigma_{ii} = 1, i = 1, \ldots, p$.

It should be noted that for each univariate continuous variable X_j, $Z_j = \Phi^{-1}(F_j(X_j))$ is standard normal where $F_j(x)$ is the cumulative distribution function of X_j and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of $N(0,1)$. This simple fact tells us that $f_j(\cdot) = \Phi^{-1}(F_j(\cdot))$ and model (1) basically assumes that after transformation those marginally normal distributed variables also follow a joint normal distribution. Of course, we cannot guarantee that marginal normal variables are jointly normal as well. Like any other semiparametric model, the semiparametric Gaussian copula model can have the model mis-specification issue when being applied in applications. However, it is clear that the semiparametric Gaussian copula model is much flexible than the normal model whilst keeping its nice interpretability. Semiparametric Gaussian copulas have generated a lot of interests in statistics, econometrics and finance [7, 16, 28, 30].

Much of the existing theoretical work on the inference of semiparametric Gaussian copulas focuses on the classical asymptotic setting where the dimension is fixed and the sample size goes to infinity. With the advances in modern technology, massive high-dimensional data are being routinely produced in various fields, including computational biology, genetics, medical imaging, climate studies, and so on. The focus of this paper is estimating $\mathbf{\Sigma}$ based on a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from model (1). To have a deep understanding of the problem, we need to address two fundamental questions.

1. What is the fundamental limit for estimating $\mathbf{\Sigma}$, when the dimension can far exceed the sample size, at least $p \ge n^\nu$ for some $\nu > 1$?
2. Can we find a data-driven method to achieve the fundamental limit?

To address these two fundamental questions we consider estimating $\mathbf{\Sigma}$ over a large parameter space stated as below

[*] Postdoctoral research associate at Princeton University. This work was finished when Lingzhou Xue was a Ph.D. student at University of Minnesota.

[†] Corresponding author. Associate Professor at University of Minnesota.

(2)

$$\begin{aligned}\mathcal{G}_q \;=\; & \big\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \text{ are random samples generated from} \\ & \text{model (1) and } \boldsymbol{\Sigma} \text{ satisfies } |\sigma_{[-j]j}|^q_{(k)} \le c_{n,p}k^{-1}, \\ & \text{for any } j=1,\ldots,p \text{ and } k=1,\ldots,p-1\big\}, \\ & 0 \le q < 1 \text{ and } c_{n,p} \le Mn^{(1-q)/2}(\log p)^{(q-3)/2},\end{aligned}$$

where $\sigma_{[-j]j}$ denotes the j-th column of $\boldsymbol{\Sigma}$ with σ_{jj} removed and the notation of $|\sigma_{[-j]j}|_{(k)}$ denotes the k-th largest element in the magnitude in $|\sigma_{[-j]j}|$. For any estimate of $\boldsymbol{\Sigma}$, denoted by $\widehat{\boldsymbol{\Sigma}}$, its quality of estimation is measured by $E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_1}$ and $E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_2}$. Note that for a matrix $\boldsymbol{A}$, its matrix ℓ_a-norm is defined as the operator norm induced by the vector ℓ_a-norm, $\|\boldsymbol{A}\|_{\ell_a} = \sup_{\|\boldsymbol{u}\|_{\ell_a}=1}\|\boldsymbol{A}\boldsymbol{u}\|_{\ell_a}$.

The following two theorems provide direct answers to questions 1 and 2, respectively.

Theorem 1. *Under the matrix ℓ_1 or ℓ_2 norm, there is a constant c such that*

$$\begin{aligned}\inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_1} \quad &\ge \quad \inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_2} \\ &\ge \quad cc^2_{n,p}\left(\log p/n\right)^{1-q}.\end{aligned}$$

Theorem 2. *An explicit rank-based thresholding estimator $\widehat{\boldsymbol{\Sigma}}^*$, which is defined in section 2.2, satisfies the risk inequality*

$$\begin{aligned}\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_2} \quad &\le \quad \sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_1} \\ &\le \quad Cc^2_{n,p}\left(\log p/n\right)^{1-q}\end{aligned}$$

for some positive constant C.

Theorems 1 and 2 imply that the estimator $\widehat{\boldsymbol{\Sigma}}^*$ is adaptive minimax optimal for estimating $\boldsymbol{\Sigma}$ under the matrix ℓ_1 and ℓ_2 norm. The minimax optimal rates of convergence for estimating sparse covariance matrices have been established in [5, 6, 35] where their parameter space $\mathcal{P}_q$ is almost $\mathcal{G}_q$ except that the distribution of the data is assumed to be sub-Gaussian and $\boldsymbol{\Sigma}$ is the population covariance matrix of the raw data. [3] showed that a data-driven adaptive thresholding estimator can attain the minimax rate of convergence. Our adaptive minimax theory shows a bigger difference between our results and previously established theory for sparse covariance matrices estimation. Note that in [3] the condition $\log p = o(n^{1/3})$ is required for establishing the adaptive minimax result. Compared with [3], our theory shows that the adaptive minimax optimal estimation is doable for ultra-high dimensions as long as $\log(p)/n \to 0$. Thus our theory can handle much higher dimensions than the adaptive minimax theory in [3].

The rest of the paper is organized as follows. Main results are presented in Section 2 where we prove Theorem 1 and Theorem 2. In Section 3 we also prove the sparsity recovery property of the rank-based thresholding when $\boldsymbol{\Sigma}$ belongs to an ℓ_0 ball. Section 4 contains numerical examples. Technical proofs are presented in an Appendix.

2. MAIN RESULTS

In this section we prove Theorems 1 and 2.

2.1 Proof of the lower bound

Because a correlation matrix can be viewed as a special covariance matrix with variance being 1, it turns out that we can directly use the lower bound results from Theorem 2 of [6] to prove the desired lower bound in our Theorem 1. In what follows we use c and C to denote generic constants in lower and upper bounds, respectively.

[6] proved Theorem 2 by considering a subspace of $\mathcal{P}_q$ denoted by $\mathcal{F}_*$ (see equation (20) in [6]), which contains a collection of normal distributions whose covariance matrices are in the weak ℓ_q ball and the diagonal elements of the covariance matrix are all 1. The readers are referred to Section 3 of [6] for the technical details. For space consideration we do not repeat these details here. It is shown in [6] that

$$(3) \qquad \inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{F}_*} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_2} \ge cc^2_{n,p}\left(\log p/n\right)^{1-q}.$$

Now we notice that $\mathcal{F}_*$ is in fact a subspace of $\mathcal{G}_q$, because each normal distribution is a special semiparametric Gaussian copula (with identity transformation) and each covariance matrix in $\mathcal{F}_*$ can be viewed as a correlation matrix since its diagonal elements are all 1. Therefore, the minimax lower bounds in Theorem 1 are proved by using (3) and the following inequalities

$$\begin{aligned}\inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_1} \quad &\ge \quad \inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_2} \\ &\ge \quad \inf_{\widehat{\boldsymbol{\Sigma}}}\sup_{\boldsymbol{\Sigma}\in\mathcal{F}_*} E\|\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|^2_{\ell_2}.\end{aligned}$$

2.2 The adaptive estimator and the upper bound

To complete the proof of Theorem 1 we now need to construct an estimator $\widehat{\boldsymbol{\Sigma}}^*$ such that

$$(4) \qquad \sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_a} \le Cc^2_{n,p}\left(\log p/n\right)^{1-q}, \quad a=1,2.$$

Note that $\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_2} \le \sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_1}$, and thus it suffices to prove that

$$(5) \qquad \sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^*-\boldsymbol{\Sigma}\|^2_{\ell_1} \le Cc^2_{n,p}\left(\log p/n\right)^{1-q}.$$

Furthermore, in order to prove Theorem 2 we need to show that the constructed estimator is fully data-driven, free of the parameter space $\mathcal{G}_q$.

A technical difficulty in constructing the estimator and proving the upper bound is how to handle these p unknown transformation functions in the semiparametric Gaussian copula model. Somewhat surprisingly, we can construct the desired estimator without estimating these transformation functions at all. Our estimator is based on the nonparametric rank estimation idea [15, 17]. Let $(x_{1i}, x_{2i}, \ldots, x_{ni})$ be the observed values of variable X_i. We convert them to ranks denoted by $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_{ni})$. Spearman's rank correlation $\hat{r}_{ij}$ is defined as Pearson's correlation between $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, Spearman's rank correlation is a nonparametric measure of dependence between two variables. Because data ranks are preserved under monotone increasing transformations, $\hat{r}_{ij}$ is also equal to the Spearman's rank correlation of the "oracle" variables Z_i, Z_j, where $Z_i = (f_i(x_{1i}), \ldots, f_i(x_{ni}))$. According to model (1) (Z_i, Z_j) follows a bivariate normal distribution with correlation parameter σ_{ij}. Then a classical result due to [15] shows that $\hat{r}^s_{ij} = 2\sin(\frac{\pi}{6}\hat{r}_{ij})$ is an asymptotically unbiased estimator σ_{ij}.

We apply the thresholding idea to obtain the desired estimator

$$\widehat{\boldsymbol{\Sigma}}^* = (s_{\lambda_*}(\hat{r}^s_{ij}))_{1\le i,j\le p},$$

where $s_\lambda(\cdot)$ applies the hard thresholding rule $h_\lambda(t) = tI(|t| > \lambda)$ to the off-diagonal elements, i.e.,

$$s_\lambda(\hat{r}^s_{ij}) = h_\lambda(\hat{r}^s_{ij}) \cdot I(i \neq j) + \hat{r}^s_{ij} \cdot I(i = j).$$

We set the thresholding parameter λ to be $\lambda_* = 40\pi \cdot (\log p/n)^{1/2}$. Note that the construction of $\widehat{\boldsymbol{\Sigma}}^*$ does not depend on parameter space $\mathcal{G}_q$. To complete the proof of Theorem 1 and Theorem 2, we only need to prove the following upper bound result

$$\sup_{\boldsymbol{\Sigma}\in\mathcal{G}_q} E\|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}\|^2_{\ell_1} \le Cc^2_{n,p}\left(\log p/n\right)^{1-q}. \tag{6}$$

The proof of (6) is given in the Appendix.

3. SPARSE RECOVERY IN THE ℓ_0 BALL CASE

If $q = 0$ we have the ℓ_0 ball case in which the correlation matrix $\boldsymbol{\Sigma}$ is strictly sparse in the sense that each row of $\boldsymbol{\Sigma}$ only has a small number of nonzero elements and the rest majority are exactly zero. [24] proved the so-called sparsistency property of the thresholding covariance estimator with sub-Gaussian data. It is interesting to note that in the semiparametric Gaussian copula model variables i and j are marginally independent if and only if $\sigma_{ij} = 0$. Recall the rank-based thresholding estimator $\widehat{\boldsymbol{\Sigma}} = (s_\lambda(\hat{r}^s_{ij}))_{1\le i,j\le p}$, and now we prove the sparsistency property of the rank-based thresholding estimator.

Theorem 3. *Let $\mathcal{A} = \{(i,j) : \ i < j,\ \sigma_{ij} \neq 0\}$ be the support set of $\boldsymbol{\Sigma}$, denoted by supp($\boldsymbol{\Sigma}$). Write $s_n = |\mathcal{A}|$ as the cardinality of $\mathcal{A}$. Define $\gamma_0 = \min_{(i,j)\in\mathcal{A}} |\sigma_{ij}|$. Pick the thresholding parameter λ satisfying that $n\gamma_0 - 12\pi \ge n\lambda \ge 12\pi$, $n\lambda^2 - 100\pi^2\log p \to \infty$ and $n(\gamma_0 - \lambda)^2 - 50\pi^2\log(s_n) \to \infty$ as $n \to \infty$. Then as $n \to \infty$, we have*

$$\begin{aligned} &\mathrm{pr}(\mathit{supp}(\widehat{\boldsymbol{\Sigma}}^*) \neq \mathit{supp}(\boldsymbol{\Sigma})) \\ \le\ & p^2\exp\left(-\frac{n\lambda^2}{50\pi^2}\right) + 2s_n\exp\left(-\frac{n(\gamma_0-\lambda)^2}{50\pi^2}\right) \to 0. \end{aligned}$$

Remark. Although the upper bound result (6) and Theorem 3 are established for the rank-based estimators using Spearman's rho, the same analysis can be easily extended to the rank-based estimators using Kendall's tau.

A related problem is to recover the sparsity pattern of $\boldsymbol{\Sigma}^{-1}$ when $\boldsymbol{\Sigma}^{-1}$ is in an ℓ_0 ball. In the semiparametric Gaussian copula model, the nonzero entries of $\boldsymbol{\Sigma}^{-1}$ correspond to the edges in a nonparametric graphical model representing the Markov dependence structure among the original variables [18, 19, 34]. [19] took a "plug-in" approach to estimate $\boldsymbol{\Sigma}^{-1}$. They first estimated $f_j(x_j)$ by $\hat{f}_j(x_j) = \Phi^{-1}(\hat{F}_j(x_j))$ where $\hat{F}_j(x_j)$ is an empirical version of $F_j(x_j)$. Then the graphical lasso estimator [9, 12, 22, 23, 37] was constructed based on the working data $(\hat{f}_1(x_{i1}), \ldots, \hat{f}_p(x_{ip}))$, $i = 1, 2, \ldots, n$. Their asymptotic theory was established for $p = O(n^\xi)$ for some $\xi > 0$. Recently, [34] and [18] independently proposed the rank-based approach for estimating $\boldsymbol{\Sigma}^{-1}$ and the theories therein work for the nearly exponentially large dimension, i.e. $\log(p) = o(n)$.

One may argue that the ℓ_0 ball case is more interesting for sparse inverse covariance (or inverse correlation) matrices because of the graphical model interpretation. Recently, [31] showed an interesting result that when computing the graphical lasso estimator of $\boldsymbol{\Sigma}^{-1}$, one can first threshold the small entries of $\boldsymbol{\Sigma}$ to zero and use those zero entries to discover the disjoint blocks in $\boldsymbol{\Sigma}^{-1}$, and thus it is sufficient to find the sparse estimates of those blocks using the graphical lasso in order to construct the graphical lasso estimator of $\boldsymbol{\Sigma}^{-1}$. In short, thresholding the sample covariance (or correlation) matrix can be used to greatly boost the computation of sparse inverse covariance (or inverse correlation) matrix estimation.

4. NUMERICAL RESULTS

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

4.1 Simulation studies

We use several simulation models to examine the finite sample performance of the proposed estimator. We first generated n independent hidden p-dimensional random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ from $N(0, \boldsymbol{\Sigma})$ and then transfer the normal data

Table 1. Simulation results of Model 1. Estimation accuracy is measured by the matrix ℓ_1-norm and ℓ_2-norm averaged over 100 replications with standard errors in parentheses

	$\rho = 0.3$			$\rho = 0.7$		
p	250	1,000	3,000	250	1,000	3,000
			Matrix ℓ_1-norm			
$\widehat{\mathbf{\Sigma}}^o$	0.93	1.13	1.22	2.39	2.70	2.84
	(0.01)	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)
$\widehat{\mathbf{\Sigma}}^*$	0.98	1.18	1.25	2.49	2.80	2.93
	(0.01)	(0.02)	(0.03)	(0.02)	(0.04)	(0.05)
			Matrix ℓ_2-norm			
$\widehat{\mathbf{\Sigma}}^o$	0.60	0.71	0.79	1.44	1.69	1.86
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)
$\widehat{\mathbf{\Sigma}}^*$	0.63	0.76	0.81	1.51	1.77	1.94
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)

Table 2. Simulation results of Model 2 & 3. Estimation accuracy is measured by both the matrix ℓ_1-norm and ℓ_2-norm averaged over 100 replications with standard errors shown in parentheses

	Model 2			Model 3		
p	250	1,000	3,000	250	1,000	3,000
			Matrix ℓ_1-norm			
$\widehat{\mathbf{\Sigma}}^o$	0.67	0.86	0.95	1.52	2.32	2.98
	(0.01)	(0.02)	(0.03)	(0.04)	(0.05)	(0.04)
$\widehat{\mathbf{\Sigma}}^*$	0.71	0.91	0.99	1.70	2.51	3.21
	(0.01)	(0.02)	(0.03)	(0.04)	(0.06)	(0.05)
			Matrix ℓ_2-norm			
$\widehat{\mathbf{\Sigma}}^o$	0.47	0.56	0.61	0.68	0.91	1.12
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
$\widehat{\mathbf{\Sigma}}^*$	0.50	0.60	0.65	0.74	0.99	1.22
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)

to the actually observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ using transformation functions in the following order

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

where $f_1(x) = x$, $f_2(x) = \log(x)$, $f_3(x) = x^{\frac{1}{3}}$ and $f_4(x) = \log(\frac{x}{1-x})$. In other words, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independent realizations from the semiparametric Gaussian copula model with the transformation functions being

$$[f_1, f_2, f_3, f_4, f_1, f_2, f_3, f_4, \ldots]$$

and the correlation matrix being $\mathbf{\Sigma}$. In our simulation we let $n = 250$ and $p = 250,\ 1{,}000\ \&\ 3{,}000$. We considered three different correlation matrices:

Model 1: $\sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.3$ and 0.7.
Model 2: $\sigma_{ij} = I_{\{i=j\}} + \rho I_{\{|i-j|=1\}}$ for $\rho = 0.3$.
Model 3: $\sigma_{ij} = s_{ij}(s_{ii}s_{jj})^{-1/2}$ where $\boldsymbol{S} = (s_{ij})_{p\times p} = (\boldsymbol{I}_{p\times p} + U)^T(\boldsymbol{I}_{p\times p} + U)$ and U is a sparse matrix with exactly p nonzero entries equal to $+1$ or -1 with equal probability.

Table 3. Simulation results of Model 2. Support recovery accuracy is measured by true positive rate / false positive rate averaged over 100 replications with standard errors in parentheses

	Model 2		
p	250	1,000	3,000
$\widehat{\mathbf{\Sigma}}^o$	0.92 / 0.00	0.84 / 0.00	0.77 / 0.00
	(0.00 / 0.00)	(0.00 / 0.00)	(0.01 / 0.00)
$\widehat{\mathbf{\Sigma}}^*$	0.88 / 0.00	0.78 / 0.00	0.70 / 0.00
	(0.00 / 0.00)	(0.00 / 0.00)	(0.01 / 0.00)

Table 4. Simulation results of Model 3. Support recovery accuracy is measured by true positive rate / false positive rate averaged over 100 replications with standard errors in parentheses

	Model 3		
p	250	1,000	3,000
$\widehat{\mathbf{\Sigma}}^o$	0.95 / 0.00	0.92 / 0.00	0.89 / 0.00
	(0.00 / 0.00)	(0.00 / 0.00)	(0.00 / 0.00)
$\widehat{\mathbf{\Sigma}}^*$	0.94 / 0.00	0.90 / 0.00	0.87 / 0.00
	(0.01 / 0.00)	(0.00 / 0.00)	(0.00 / 0.00)

These models have been used in previous works [2, 24, 26, 27]. Model 1 belongs to the weak ℓ_q ball case, and Model 2 and 3 are the ℓ_0 ball case. Model 3 has a random ℓ_0 ball structure, and on average it has 990, 3,981 and 12,008 nonzero entries for $p = 250, 1{,}000$ and $3{,}000$ respectively.

The goal is to use simulation to confirm the theoretical finding. To this end, we include the "oracle" estimator as the benchmark in our simulation study. The "oracle" estimator is constructed by thresholding the sample correlation matrix of the hidden data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, because the oracle knows the true transformation functions. Our rank-based estimator is constructed using the observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We used the 5-fold cross validation [2, 3, 24] to tune both estimators. For ease of notation, we denote the "oracle" estimator by $\widehat{\mathbf{\Sigma}}^o$ and the proposed rank-based estimator by $\widehat{\mathbf{\Sigma}}^*$.

The simulation results are summarized in Tables 1–4. For all three simulation models, we compare the estimation performance using both the matrix ℓ_1-norm and the matrix ℓ_2-norm. In the simulation models 2 and 3, we also report the sparsity recovery performance using the true positive rate and the false positive rate [24], i.e.,

$$\frac{\#\{(i,j):\ \hat{\sigma}_{ij} \neq 0,\ \&\ \sigma_{ij} \neq 0\}}{\#\{(i,j):\ \sigma_{ij} \neq 0\}}$$

and

$$\frac{\#\{(i,j):\ \hat{\sigma}_{ij} \neq 0,\ \&\ \sigma_{ij} = 0\}}{\#\{(i,j):\ \sigma_{ij} = 0\}}.$$

From Tables 1–3, the proposed rank estimator works as well as the "oracle" estimator, which is what our theory predicts.

Table 5. Normality test results for the Arcene data. The counts of genes that fail to pass each normality test are shown in the table

Data	Cancer		Healthy	
Critical value	0.01	0.01/200	0.01	0.01/200
Anderson–Darling	197	190	196	188
Lilliefors	194	183	194	157
Pearson's χ^2	198	188	194	165
Shapiro–Francia	197	180	195	172

The "oracle" estimator is slightly better than the rank estimator, which is expected because some information is lost when converting the original data into ranks.

In our simulation study, we also tried the rank-based estimators using Kendall's tau. The corresponding simulation results are nearly identical to that of the rank estimators using Spearman's rho.

4.2 Arcene data

We use the Arcene mass-spectrometric data [14] to demonstrate the use of the semiparametric Gaussian copula model and the proposed estimator. This dataset includes 200 samples with 112 healthy patients and 88 cancer patients with ovarian or prostate tumors from the National Cancer Institute and the Eastern Virginia Medical School. Each sample has 7,000 original features indicating the abundance of proteins in human sera having a given mass value, and 3,000 distractor features having no predictive power. The Arcene dataset can be downloaded from the UCI Machine Learning Repository [11].

We first performed the Kolmogorov filter [20] to pick the top 200 features. The Kolmogorov filter is a fully nonparametric screening method for the data with binary responses, and this method is shown to enjoy the sure screening property under weak assumptions [20]. Denote by $\hat{F}^h_j(x)$ and $\hat{F}^c_j(x)$ the empirical conditional cumulative distribution function of the j-th feature for healthy patients and cancer patients, respectively. The Kolmogorov filter ranks all features by the corresponding two-sample Kolmogorov-Smirnov test statistic $\hat{K}_j = \sup_{-\infty<x<+\infty} |\hat{F}^h_j(x) - \hat{F}^c_j(x)|$, and selects 200 features whose $\hat{K}_j$ are amongst the first 200 largest of all $\hat{K}_j$'s. We then conducted various normality tests on these top 200 features. The test results are shown in Table 5. For both cancer and healthy patients, more than 90% genes are unable to pass any of four normality tests at the significance level of 0.01, and under Bonferroni correction there are still over 75% genes that fail to pass the normality tests. Figure 1 plotted the histograms of the mass-spectrometric values for the top 4 features in terms of the Kolmogorov-Smirnov statistic in the Kolmogorov filter. From Figure 1 we see that the features can have a bimodal distribution or a highly skewed empirical distribution for both cancer and healthy patients.

To deal with the non-normal issue, we consider the semiparametric Gaussian copula model to this dataset. In the Arcene data, the order of features were randomized [14]. We are interested in comparing the correlation matrices between the healthy patient group and the cancer patient group. Because there is no reliable ordering information, we applied the rank-based thresholding estimator to these 200 features for cancer and healthy patients respectively. Figure 2 shows the heatmaps of the corresponding rank-based thresholding estimators. The features in Panel (A) and (B1) are ordered by hierarchical clustering using the estimated correlations for healthy and cancer patients respectively. We also plotted the heatmap for the cancer patients according to the order by hierarchical clustering using the estimated correlations of healthy patients in Panel (B2). Both Panel (B1) and Panel (B2) have shown a quite different pattern from Panel (A). We further calculated the statistic $L_2 = \|\widehat{\boldsymbol{\Sigma}}^*_{\text{cancer}} - \widehat{\boldsymbol{\Sigma}}^*_{\text{healthy}}\|_{\ell_2} = 41.49$. We further computed the 95% bootstrap confidence interval of L_2 based on $B = 500$ bootstrapped random samples. The bootstrap confidence interval is $[30.51, \ 53.26]$, which clearly indicates that the correlation structures among two groups are different.

5. DISCUSSION

Besides thresholding, there are several other useful regularization techniques for high-dimensional covariance estimation, such as banding [1, 32], tapering [4, 13], Cholesky-based regularization [1, 25] and positive definite ℓ_1 penalized estimation [33]. These techniques can be combined with the rank-based correlation estimation idea for estimating the correlation matrices of semiparametric Gaussian copulas, if $\boldsymbol{\Sigma}$ is assumed to have a different structure that is more suitable for applying these techniques. [36] studied rank-based tapering estimator. In this work we focus on thresholding estimation because it is permutation invariant, which is a big advantage over banding/tapering and Cholesky-based regularization when there is no reliable ordering information about the variables [2, 24].

APPENDIX A. TECHNICAL PROOFS

Our proof uses the following useful concentration bound whose proof is given in [34].

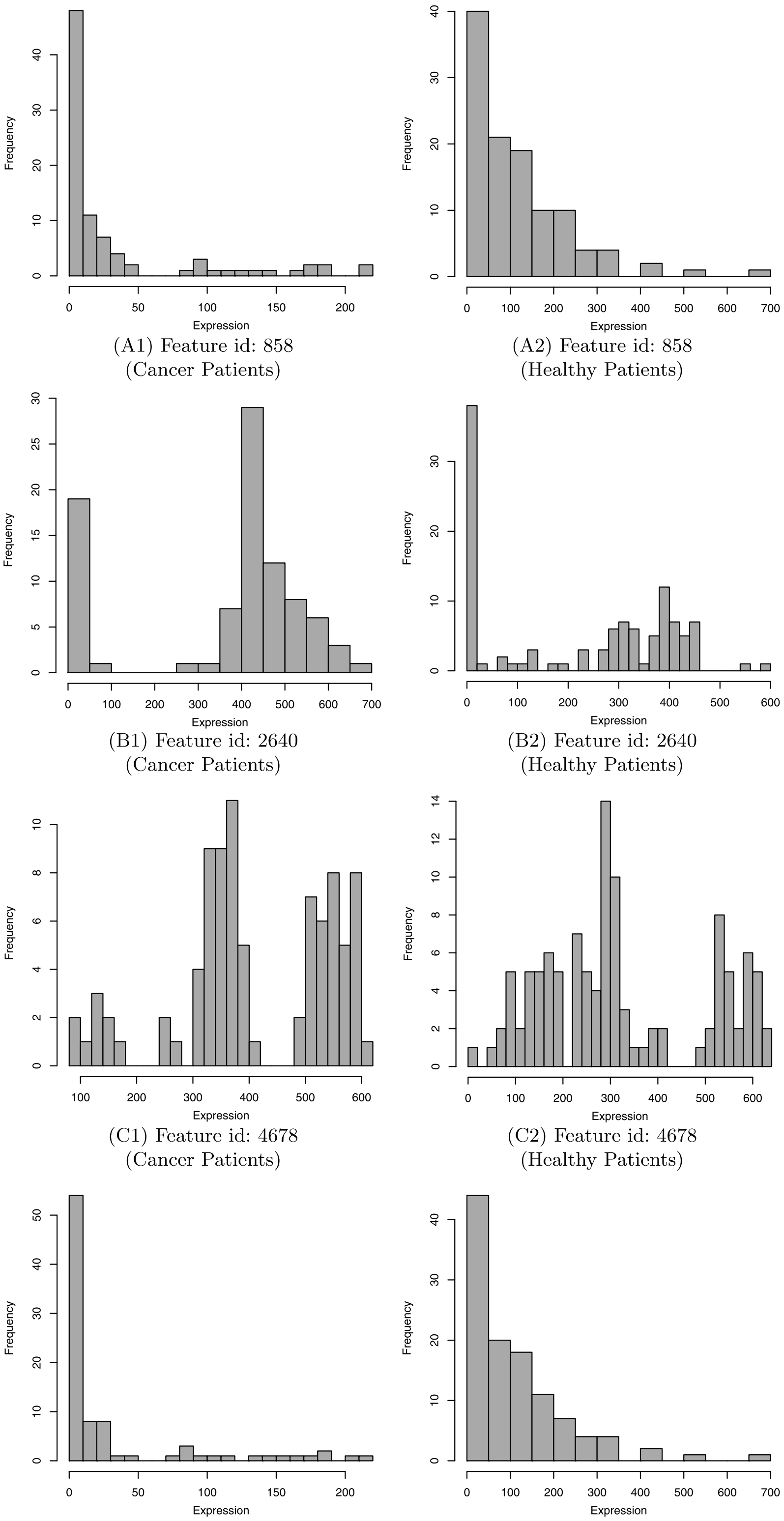
Lemma 1. *Fix any $0 < \varepsilon < 1$ and suppose that $n\varepsilon \geq 12\pi$, we have*

$$\Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \varepsilon) \leq 2 \exp\left(- \frac{n\varepsilon^2}{50\pi^2} \right).$$

A.1 Proof of Equation (6)

As we discussed in Section 2, the proof of Theorem 1 is boiled down to the proof of (6). First we derive the probability upper bound for $|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}|$. Notice that

$$|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}| = |\sigma_{ij}| \cdot I_{\{|\hat{r}^s_{ij}| \leq \lambda_*\}} + |\hat{r}^s_{ij} - \sigma_{ij}| \cdot I_{\{|\hat{r}^s_{ij}| > \lambda_*\}}.$$



(A1) Feature id: 858
(Cancer Patients)

(A2) Feature id: 858
(Healthy Patients)

(B1) Feature id: 2640
(Cancer Patients)

(B2) Feature id: 2640
(Healthy Patients)

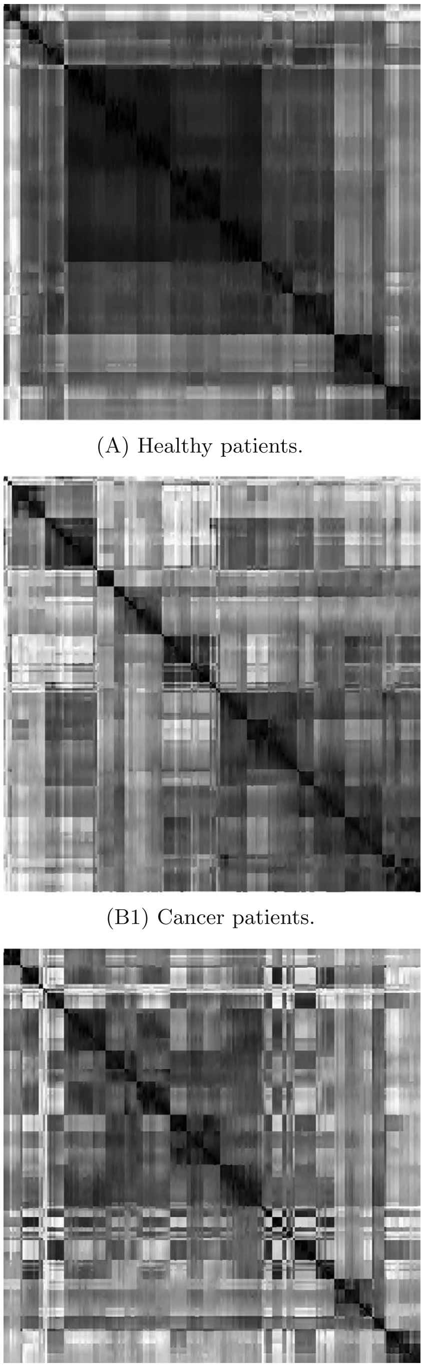
(C1) Feature id: 4678
(Cancer Patients)

(C2) Feature id: 4678
(Healthy Patients)

(D1) Feature id: 6442
(Cancer Patients)

(D2) Feature id: 6442
(Healthy Patients)

Figure 1. Illustration of non-normality using histograms of the mass-spectrometric values in the Arcene data.



(A) Healthy patients.

(B1) Cancer patients.

(B2) Cancer patients.

Figure 2. Heapmaps of the absolute values of thresholding estimators for the Arcene data. The features are ordered by hierarchical clustering using the estimated correlations.

Following [5] we consider the following three possible settings with respect to $|\sigma_{ij}|$.

(i) when $|\sigma_{ij}| < \frac{1}{2}\lambda_*$, $\{|\hat{r}^s_{ij}| > \lambda_*\} \subset \{|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*\}$ obviously holds by the triangle inequality $|\hat{r}^s_{ij} - \sigma_{ij}| > |\hat{r}^s_{ij}| - |\sigma_{ij}| > \frac{1}{2}\lambda_*$. Then we have $|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}| = |\sigma_{ij}|$ with probability at least $1 - \Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*) = 1 - 2p^{-8}$.

(ii) when $\frac{1}{2}\lambda_* \leq |\sigma_{ij}| \leq \frac{3}{2}\lambda_*$, $|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}| \leq \max\{|\sigma_{ij}|, |\hat{r}^s_{ij} - \sigma_{ij}|\} \leq |\sigma_{ij}|$ with probability of at least $1 - \Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*) = 1 - 2p^{-8}$.

(iii) when $|\sigma_{ij}| > \frac{3}{2}\lambda_*$, we have $\{|\hat{r}^s_{ij}| \leq \lambda_*\} \subset \{|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*\}$ by the triangle inequality $|\hat{r}^s_{ij} - \sigma_{ij}| > |\sigma_{ij}| - |\hat{r}^s_{ij}| > \frac{1}{2}\lambda_*$. Then we have $|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}| = |\hat{r}^s_{ij} - \sigma_{ij}| \leq \frac{1}{2}\lambda_*$ with probability at least $1 - \Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*) = 1 - 2p^{-8}$.

For all three scenarios, we always have $|s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}| \leq \min\{|\sigma_{ij}|, \frac{3}{2}\lambda_*\}$ with a high probability of at least $1 - \Pr(|\hat{r}^s_{ij} - \sigma_{ij}| > \frac{1}{2}\lambda_*) = 1 - 2p^{-8}$.

Pick $k^* = \lfloor c_{n,p}(\log p/n)^{-q/2} \rfloor$, and thus

$$(c_{n,p}/k^*)^{1/q} \geq (\log p/n)^{1/2} \geq [c_{n,p}/(k^*+1)]^{1/q}.$$

Uniformly for any $i = 1, \ldots, p$ we have that,

$$\begin{aligned} & \sum_{j=1}^{p} \min\left\{|\sigma_{ij}|, \frac{3}{2}\lambda_*\right\} \cdot I(j \neq i) \\ \leq\ & \frac{3}{2}\lambda_* \cdot k^* + \sum_{i > k^*} \left(\frac{c_{n,p}}{i}\right)^{1/q} \\ \leq\ & C c_{n,p} {\lambda_*}^{1-q} + C c_{n,p}^{1/q} (k^*)^{1-1/q} \\ \leq\ & C c_{n,p} {\lambda_*}^{1-q}. \end{aligned}$$

Then we can derive the desired upper bound as follows

$$\begin{aligned} & \mathbb{E}\left\|\widehat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}\right\|_{\ell_1}^2 \\ =\ & \mathbb{E}\left(\max_{i=1,\ldots,p} \sum_{j=1}^{p} |s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}|\right)^2 \\ & \times I\left(\max_{(i,j)} |\hat{r}^s_{ij} - \sigma_{ij}| \leq \min\left\{|\sigma_{ij}|, \frac{3}{2}\lambda_*\right\}\right) \\ & + \mathbb{E}\left(\max_{i=1,\ldots,p} \sum_{j=1}^{p} |s_{\lambda_*}(\hat{r}^s_{ij}) - \sigma_{ij}|\right)^2 \\ & \times I\left(\max_{(i,j)} |\hat{r}^s_{ij} - \sigma_{ij}| > \min\left\{|\sigma_{ij}|, \frac{3}{2}\lambda_*\right\}\right) \\ \leq\ & E\left(\sum_{j=1}^{p} \min\left\{|\sigma_{ij}|, \frac{3}{2}\lambda_*\right\} \cdot I(j \neq i)\right)^2 \\ & + 4p^2 \Pr\left\{\max_{(i,j)} |\hat{r}^s_{ij} - \sigma_{ij}| > \min\left\{|\sigma_{ij}|, \tfrac{3}{2}\lambda_*\right\}\right\} \\ \leq\ & C c_{n,p}^2 {\lambda_*}^{2-2q} + C p^4 \cdot p^{-8} \\ \leq\ & C c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}. \end{aligned}$$

This completes the proof of Equation (6).

A.2 Proof of Theorem 3

First we notice that

$$\begin{aligned} & \{\mathrm{supp}(\widehat{\mathbf{\Sigma}}^*) \neq \mathrm{supp}(\mathbf{\Sigma})\} \\ \subseteq \; & \Big\{\max_{(i,j)\in\mathcal{A}^c} |\hat{r}^s_{ij}| > \lambda\Big\} \cup \Big\{\max_{(i,j)\in\mathcal{A}} |\hat{r}^s_{ij}| \leq \lambda\Big\} \\ \subseteq \; & \Big\{\max_{(i,j)\in\mathcal{A}^c} |\hat{r}^s_{ij} - \sigma_{ij}| > \lambda\Big\} \\ & \cup \Big\{\max_{(i,j)\in\mathcal{A}} |\hat{r}^s_{ij} - \sigma_{ij}| > \gamma_0 - \lambda\Big\}. \end{aligned}$$

Then by using the concentration bound for $\hat{r}^s_{ij} - \sigma_{ij}$ from Lemma 1, we have

$$\begin{aligned} & \mathrm{pr}(\mathrm{supp}(\widehat{\mathbf{\Sigma}}^*) \neq \mathrm{supp}(\mathbf{\Sigma})) \\ \leq \; & \mathrm{pr}\Big(\max_{(i,j)\in\mathcal{A}^c} |\hat{r}^s_{ij} - \sigma_{ij}| > \lambda\Big) \\ & + \mathrm{pr}\Big(\max_{(i,j)\in\mathcal{A}} |\hat{r}^s_{ij} - \sigma_{ij}| > \gamma_0 - \lambda\Big) \\ \leq \; & p^2 \exp(-c_0 n \lambda^2) + 2 s_n \exp(-c_0 n (\gamma_0 - \lambda)^2). \end{aligned}$$

This completes the proof of Theorem 3.

ACKNOWLEDGEMENT

The authors thank the editor, associate editor and referees for their helpful comments. This work is partly supported by a grant from the National Science Foundation and a grant from the Office of Naval Research.

Received 22 March 2013

REFERENCES

[1] Bickel, P. & Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36**, 199–227. MR2387969

[2] Bickel, P. & Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604. MR2485008

[3] Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607. MR2847973

[4] Cai, T., Zhang, C. & Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118–2144. MR2676885

[5] Cai, T. & Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm (with discussion). *Statistica Sinica* **22**, 1319–1378. MR3027084

[6] Cai, T. & Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* **40**, 2389–2420. MR3097607

[7] Chen, X. & Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics* **130**, 307–335. MR2211797

[8] Chen, X., Fan, Y. & Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association* **101**, 1228–1240. MR2328309

[9] d'Aspremont, A., Banerjee, O. & El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications* **30**, 56–66. MR2399568

[10] El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *The Annals of Statistics* **36**, 2717–2756. MR2485011

[11] Frank, A. & Asuncion, A. (2010). Uci machine learning repository [http://archive.ics.uci.edu/ml/]. University of California, School of Information and Computer Science, Irvine, CA.

[12] Friedman, J., Hastie, T. & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

[13] Furrer, R. & Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* **98**, 227–255. MR2301751

[14] Guyon, I., Gunn, S., Ben-Hur, A. & Dror, G. (2004). Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems* **17**, 545–552.

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

[16] Klaassen, C. & Wellner, J. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3**, 55–77. MR1466545

[17] Lehmann, E. (1998). *Nonparametrics: Statistical Methods Based on Ranks.* Prentice-Hall, New Jersey. MR2279708

[18] Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). High dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics* **40**, 2293–2326. MR3059084

[19] Liu, H., Lafferty, J. & Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**, 1–37. MR2563983

[20] Mai, Q. & Zou, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100**, 229–234. MR3034336

[27] Ma, S., Xue, L. & Zou, H. (2013). Alternating direction methods for latent variable Gaussian graphical model selection. *Neural computation* **25**, 2172–2198. MR3100000

[22] Ravikumar, P., Wainwright, M., Raskutti, G. & Yu, B. (2008). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Advances in Neural Information Processing Systems.*

[23] Rothman, A., Bickel, P., Levina, E. & Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515. MR2417391

[24] Rothman, A., Levina, E. & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177–186. MR2504372

[25] Rothman, A., Levina, E. & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97**, 539–550. MR2672482

[26] Scheinberg, K., Ma, S. & Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In *Proceedings of the Neural Information Processing Systems (NIPS).*

[27] Rothman, A., Levina, E. & Zhu, J. (2010). Alternating direction methods for latent variable Gaussian graphical model selection. *Biometrika* **25**, 2172–2198. MR2672482

[28] Song, P. (2000). Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics* **27**, 305–320. MR1777506

[29] Song, P., Li, M. & Yuan, Y. (2000). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65**, 60–68. MR2665846

[30] Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canadian Journal of Statistics* **33**, 357–375. MR2193980

[31] Witten, D., Friedman, J. & Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics* **20**, 892–900. MR2878953

[32] Wu, W. & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831–844. MR2024760

[33] Xue, L., Ma, S. & Zou, H. (2012). Positive definite ℓ_1 penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, **107**, 1480–1491. MR3036409

[34] Xue, L. & Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, **40**, 2541–2571. MR3097612

[35] Xue, L. & Zou, H. (2013). Minimax optimal estimation of general bandable covariance matrices. *Journal of Multivariate Analysis*, **116**, 45–51. MR3049890

[36] Xue, L. & Zou, H. (2014). Rank-based tapering estimation of bandable correlation matrices. *Statistica Sinica* **24**, in press.

[37] Yuan, M. & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35. MR2367824

Lingzhou Xue
Department of Operations Research
and Financial Engineering
Princeton Univeristy
Princeton, NJ 08544
USA
E-mail address: lzxue@princeton.edu

Hui Zou
School of Statistics
University of Minnesota
Minneapolis, MN 55455
USA
E-mail address: zouxx019@umn.edu