Gaussian Copula Precision Estimation with Missing Values

Huahua Wang **Faridel Fazayeli** **Soumyadeep Chatterjee** **Arindam Banerjee**
Department of Computer Science and Engineering, University of Minnesota, Twin cities

Abstract

We consider the problem of estimating sparse precision matrix of Gaussian copula distributions using samples with missing values in high dimensions. Existing approaches, primarily designed for Gaussian distributions, suggest using plugin estimators by disregarding the missing values. In this paper, we propose double plugin Gaussian (DoPinG) copula estimators to estimate the sparse precision matrix corresponding to *non-paranormal* distributions. DoPinG uses two plugin procedures and consists of three steps: (1) estimate nonparametric correlations based on observed values, including Kendall's tau and Spearman's rho; (2) estimate the nonparanormal correlation matrix; (3) plug into existing sparse precision estimators. We prove that DoPinG copula estimators consistently estimate the non-paranormal correlation matrix at a rate of $O(\frac{1}{(1-\delta)}\sqrt{\frac{\log p}{n}})$, where δ is the probability of missing values. We provide experimental results to illustrate the effect of sample size and percentage of missing data on the model performance. Experimental results show that DoPinG is significantly better than estimators like mGlasso, which are primarily designed for Gaussian data.

1 Introduction

In recent years, considerable effort [1, 6, 17, 18, 4, 3, 14, 25] has been invested in obtaining an accurate estimate of the precision matrix based on the sample covariance matrix, especially when the true precision matrix is assumed to be sparse [25]. Suitable estimators and corresponding statistical convergence rates have been established for a variety of settings, including distributions with sub-Gaussian tails, polynomial tails [18, 4, 14].

Although these sparse precision estimators are primarily designed to deal with fully observed data, recently, they have also been generalized to handle data with missing values [11, 19, 16, 15, 9], which often occur in real world applications, e.g., drop-outs of sensors in a sensor network or missing measurements of temperature or rain in climate. To deal with data with missing values, a variety of methods apply expectation maximization (EM) algorithms on imputed data, which are iterative methods but lack theoretical guarantees [11, 19]. In particular, [19] proposed an EM algorithm named MissGlasso to deal with missing values using Glasso. MissGlasso first imputes the missing values in the E-step and then solves the Glasso problem on the imputed data in the M-step. As EM converges to a local optimum, it is difficult to establish theoretical guarantees for the MissGlasso procedure. Without using the EM algorithm, [15] employed projected gradient descent to solve a sequence of regression problems or PGlasso to estimate the sparse precision matrix of incomplete data. Theoretical guarantees are also established for the PGlasso estimator. [9] introduced a simple plug-in procedure for incomplete data which simply applies existing estimators to the observed data by disregarding the missing values. Such simple plug-in estimators for missing values can leverage existing theoretical results and thus still have similar statistical guarantees, including rate of convergence and consistency. However, these sparse precision estimators rely on the Gaussian assumption, which may not be appropriate for real datasets which are usually non-Gaussian.

To deal with non-Gaussian data, [12] proposed Gaussian copula graphical models where existing estimators can be generalized to the *non-paranormal* distributions simply using one additional procedure, i.e., estimating nonparametric correlations. Non-paranormal distributions can be considered as a non-parametric extension of the normal distribution where suitable univariate monotone transformations of the covariates are jointly distributed as a multivariate Gaussian. It has also been shown that the nonparanormal is equivalent to Gaussian copula distribution [13, 21, 20]. Therefore, the estimated correlation matrix of the data after transformation can be plugged into the standard sparse precision estimators with Gaussian assumption. The plug-in procedure can leverage existing theoretical results and achieve the optimal statistical rate of convergence. A similar procedure has also been studied independently by [24].

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

However, whether Gaussian copula graphical models can deal with missing values and maintain the optimal statistical rate of convergence is still unknown.

In this paper, we propose Double Plug-in Gaussian (DoPinG) copula estimators to deal with missing values, which estimates the sparse precision matrix corresponding to the non-paranormal distribution. DoPingG copula estimators essentially combines two plug-in procedures for dealing with missing values [9] and non-Gaussian data [12], yielding a fairly rich family of estimators to deal with incomplete data from the non-paranormal family. Such estimators consider the following three steps: (1) estimate non-parametric correlations, such as Kendall's tau and Spearman's rho, between all pairs of covariates by suitably disregarding missing values; (2) estimate the non-paranormal correlation matrix using the Kendall's tau or Spearman's rho correlation matrix; (3) plug the estimated correlation matrix into existing sparse precision estimators, e.g., graphical LASSO [1, 6], Dantzig selector [25], CLIME [4], etc.

Our analysis follows the development in [12] with one important difference: the samples we consider can have missing values. We investigate how missing values affect the accuracy of covariance estimation, and in turn precision estimation. In particular, the theoretical analysis of DoPinG copula estimators considers two probability spaces, i.e., probability over samples and probability over missing values. We assume that the data is missing completely at random (MCAR) [9], where any element is missing with probability δ. We prove that DoPinG copula estimators consistently estimate the non-paranormal correlation matrix at a rate of $O(\frac{1}{(1-\delta)}\sqrt{\frac{\log p}{n}})$.

For estimating the precision matrix, one can use any of the available estimators, such as the graphical lasso [1], graphical Dantzig selector [25], as discussed in [12, 9]. We consider the CLIME estimator [4] for our analysis. The CLIME estimator has strong statistical guarantees for consistency along with rates [4], and also comes with inherent computational advantages [23]. In particular, a large scale distributed algorithm has been developed in [23], which can scale up to millions of dimensions and trillions of parameters, using hundreds of cores. We provide experimental results to show the effect of sample size and percentage of missing data on the model performance. Experimental results show that DoPinG is significantly better than estimators like mGlasso, which are primarily designed for Gaussian data.

The rest of paper is organized as follows. We propose non-paranormal dual plug-in estimators with missing values in Section 2. In Section 3, we give the theoretical guarantees in terms of rates of convergences under element-wise L_∞ norm. We present experimental results in Section 4, and conclude the paper in Section 5.

2 Gaussian Copula Precision Estimation with Missing Values

We consider a p-dimensional *non-paranormal* distribution [12]. For univariate monotone functions $f_1, \ldots, f_p$ and a positive definite correlation matrix $\Sigma^0 \in \mathbb{R}^{p\times p}$, a p-dimensional random variable $X = (X_1, \ldots, X_p)^T$ has a non-paranormal distribution $X \sim \text{NPN}_p(f, \Sigma^0)$ if $f(X) = (f_1(X_1), \ldots, f_p(X_p)) \sim N_p(0, \Sigma^0)$, a p-dimensional multi-variate Gaussian distribution with correlation matrix Σ^0. We focus on estimating the sparse precision matrix $\Omega_0 = \Sigma_0^{-1}$ corresponding to the non-paranormal distribution.

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be samples drawn independently from $\text{NPN}_p(f, \Sigma^0)$. We further assume that for dimension j, x_{ij} will be missing with probability $\delta \in [0, 1]$. Let $b_{ij} = 1$ if x_{ij} is observed, and $b_{ij} = 0$ otherwise. Thus, $P(b_{ij} = 1) = 1 - \delta$. We assume the data is missing completely at random (MCAR) [9].

In order to estimate the precision matrix Ω^0 using CLIME, we need an empirical estimate $\hat{S}_n$ of the correlation matrix Σ^0. In particular, the elementwise L_∞ norm between the matrices need to be suitably bounded for norm consistency of precision estimation. As shown in [12] , $\hat{S}_n$ can be efficiently computed from the empirical Kendall's tau or Spearman's rho correlation matrix. Hereafter, for ease of notation, we drop the subscript n on $\hat{S}$ and other sample estimates.

DoPinG copula estimators consider three steps in estimating the precision matrix. First, suitably generalizing the plug-in procedure for estimating non-parametric correlations to handle missing values, pairwise Kendall's tau or Spearman's rho correlation between covariates is estimated. Second, the correlation matrix corresponding to the non-paranormal distribution is estimated using the Kendall's tau or Spearman's rho correlation matrices. Third, the precision matrix is estimated by simply plugging in the estimated correlation matrix into existing sparse precision matrix estimators. We discuss each one of these steps below.

2.1 Kendall's tau with missing values

Given that samples have missing values, we compute the Kendall's tau for dimensions (j, k) using the n_{jk} *effective* independent samples which have values for both dimensions. In particular, we estimate Kendall's rho as:

$$\hat{\tau}_{jk} = \frac{1}{n_{jk}(n_{jk}-1)} \sum_{\substack{i,i'=1 \\ i\neq i'}}^{n} b_{ij} b_{ik} b_{i'j} b_{i'k} \text{sign}((x_i^j - x_{i'}^j)(x_i^k - x_{i'}^k)) \ , \tag{1}$$

where $n_{jk} = \sum_{i=1}^n b_{ij} b_{ik}$. Note for the i-th sample, both the j- and k-th dimensions should not be missing. In other

words, the samples with missing values will not be considered in the estimation of the Kendall' tau.

The second step is to estimate the correlation matrix directly based on the Kendall's tau. Following [12, 10, 5], we consider the following estimator $\hat{S}^\tau = [\hat{S}^\tau_{jk}]$ for the estimated correlation matrix Σ^0:

$$\hat{S}^\tau_{jk} = \begin{cases} \sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if } j = k\,. \end{cases} \tag{2}$$

2.2 Spearman's rho with missing values

Similar to the estimation of Kendall's tau for missing values, we also compute the Spearman's rho for dimensions (j,k) using the n_{jk} *effective* independent samples which have values for both dimensions. In particular, $n_{jk} = \sum_{i=1}^n b_{ij}b_{ik}$. Let r_i^j be the rank of x_i^j among the n_{jk} samples with values and $\bar{r}_{jk}$ be the average, i.e., $\bar{r}_{jk} = \frac{1}{n_{jk}}\sum_{i=1}^n r_i^j b_{ij}b_{ik}$. Spearman's rho is defined as follows:

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^n (r_i^j - \bar{r}_{jk})(r_i^k - \bar{r}_{jk}) b_{ij}b_{ik}}{\sqrt{\sum_{i=1}^n [(r_i^j - \bar{r}_{jk})^2 b_{ij}b_{ik}] \sum_{i=1}^n [(r_i^k - \bar{r}_{jk})^2 b_{ij}b_{ik}]}}\,, \tag{3}$$

which is the first step in DoPinG.

Based on the estimate of the Spearman's rho (3), following [12, 24] , the second step is to estimate $\hat{S}^\rho = [\hat{S}^\rho_{jk}]$ for the unknown correlation matrix Σ^0:

$$\hat{S}^\rho_{jk} = \begin{cases} 2\sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right) & \text{if } j \neq k \\ 1 & \text{if } j = k\,. \end{cases} \tag{4}$$

2.3 Plugin estimate for CLIME

Having obtained $\hat{S}$ ($\hat{S}^\tau$ or $\hat{S}^\rho$), we can plugin it into any sparse precision estimators, e.g., graphical lasso [1], graphical Dantzig selector [25], CLIME [4]. In particular, we plugin $\hat{S}$ into the CLIME estimator [24]:

$$\hat{\Omega}_n = \text{argmin}_{\hat{\Omega}} \|\hat{\Omega}\|_1 \quad \text{s.t.} \quad \|\hat{S}\hat{\Omega} - \mathbf{I}\|_\infty \leq \lambda_n\,, \tag{5}$$

where λ_n is a tuning parameter and $\mathbf{I}$ is an identity matrix. The CLIME estimator has strong statistical guarantees [4], and also comes with inherent computational advantages. The estimator can scale up to millions of dimensions and can be run on hundreds of cores [23]. In [23], (5) is decomposed into solving $\lceil p/k \rceil$ independent column block linear programs where each column block contains $k (1 \leq k \leq p)$ columns. Denoting $\mathbf{X} \in \Re^{p\times k}$ be k columns of $\hat{\Omega}$, (5) can be written as

$$\min \|\mathbf{P}\|_1 \quad \text{s.t.} \quad \|\hat{S}\mathbf{P} - \mathbf{E}\|_\infty \leq \lambda_n\,, \tag{6}$$

which can be solved by an inexact ADMM algorithm [2, 22] given in Algorithm 1 [23] where ρ, η are parameters of ADMM and

Algorithm 1 Column Block Inexact ADMM for CLIME

1: **Input:** $\hat{S}, \lambda_n, \rho, \eta$
2: **Output:** $\mathbf{P}$
3: **Initialization:** $\mathbf{P}^0, \mathbf{Z}^0, \mathbf{Y}^0, \mathbf{V}^0, \hat{\mathbf{V}}^0 = 0$
4: **for** $t = 0$ to $T-1$ **do**
5: **X-update:** $\mathbf{P}^{t+1} = \mathit{soft}(\mathbf{P}^t - \mathbf{V}^t, \frac{1}{\eta})$, where
6: **Mat-Mul:** $\mathbf{U}^{t+1} = \hat{S}\mathbf{P}^{t+1}$
7: **Z-update:** $\mathbf{Z}^{t+1} = \mathit{box}(\mathbf{U}^{t+1} + \mathbf{Y}^t, \lambda_n)$, where
8: **Y-update:** $\mathbf{Y}^{t+1} = \mathbf{Y}^t + \mathbf{U}^{t+1} - \mathbf{Z}^{t+1}$
9: **Mat-Mul:** $\hat{\mathbf{V}}^{t+1} = \hat{S}\mathbf{Y}^{t+1}$
10: **V-update:** $\mathbf{V}^{t+1} = \frac{\rho}{\eta}(2\hat{\mathbf{V}}^{t+1} - \hat{\mathbf{V}}^t)$
11: **end for**

$$\mathit{soft}(\mathbf{P}, \gamma) = \begin{cases} P_{ij} - \gamma\,, & \text{if } P_{ij} > \gamma\,, \\ P_{ij} + \gamma\,, & \text{if } P_{ij} < -\gamma\,, \\ 0\,, & \text{otherwise} \end{cases}$$

$$\mathit{box}(\mathbf{P}, \mathbf{E}, \lambda_n) = \begin{cases} E_{ij} + \lambda\,, & \text{if } P_{ij} - E_{ij} > \lambda_n\,, \\ P_{ij}\,, & \text{if } |P_{ij} - E_{ij}| \leq \lambda_n\,, \\ E_{ij} - \lambda\,, & \text{if } P_{ij} - E_{ij} < -\lambda_n\,, \end{cases}$$

While steps 5, 7, 8 and 10 amount to elementwise operations, the most intensive computation is matrix multiplication in steps 6 and 9 which can be solved in parallel.

Note that the estimated correlation matrix $\hat{S}$ ($\hat{S}^\tau$ or $\hat{S}^\rho$) may be not positive semi-definite. Sparse precision estimators do require the positive semi-definiteness assumption in theory and most algorithms may fail if the input correlation matrix is not positive semi-definite [12, 9]. The inexact ADMM algorithm for CLIME in Algorithm 1 does not necessarily require $\hat{S}$ to be positive semi-definite. As long as the linear programs (5) have solutions, Algorithm 1 still works, although there is no guarantee that the solution is positive definite. Therefore, one may project the input correlation matrix onto the cone of positive semi-definite matrix in order to obtain a positive definite precision matrix with high probability using Algorithm 1. We study the effect of the two choices on the performance of DoPinG in experiments in Section 4.

3 Theoretical Analysis

In this section, we present statistical guarantees for the proposed DoPinG by leveraging existing analysis in [12, 4, 24]. Note that the consistency analysis of the CLIME estimate $\hat{\Omega}$ relies on obtaining a consistent estimate of the covariance Σ^0, defined in terms of the elementwise L_∞ norm of the difference $(\hat{S} - \Sigma^0)$. Therefore, we first analyze $\sup_{jk} \left|\hat{S}^\tau_{jk} - \Sigma^0_{jk}\right|$ for the Kendall's tau ($\hat{S} = \hat{S}^\tau$) and Spearman's rho ($\hat{S} = \hat{S}^\rho$) seperately. Our proof operates on two probability spaces, i.e., probabilities over the samples $\mathbb{P}_X$ and probabilities over the Bernoulli missing values $\mathbb{P}_B$. Then, we plug the results into the consistency

analysis of the CLIME to obtain the optimal statistical rate of convergence.

We first consider the probabilities over missing values in the following lemma which we need in the analysis of Kendall's tau and Spearman's rho:

Lemma 3.1 *Let $B = [b_{ij}] \in \{0,1\}^{n\times p}$ be an binary matrix. Assume b_{ij} is i.i.d. with a Bernoulli distribution where $P(b_{ij} = 0) = \delta$ and $P(b_{ij} = 1) = 1 - \delta$. Let $n_{jk} = \sum_{i=1}^{n} b_{ij}b_{ik}$. For any $m > 0$, and any $0 < \epsilon < 1$, we have*

$$\mathbb{P}_B\left(\sum_{j,k}\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right) > \frac{1}{p^m}\right\} \leq \exp\left(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p)\right) , \quad (7)$$

Proof: Since n_{jk} is a sum of n independent Bernoulli random variables $b_{ij}b_{ik}$ with $P(b_{ij}b_{ik} = 1) = (1-\delta)^2$, by linearity of expectation and independence of samples, we have $E[n_{jk}] = \sum_{i=1}^{n} E[b_{ij}b_{ik}] = n(1-\delta)^2$. By standard Chernoff bounds, for any $\epsilon < 1$, we have

$$\begin{aligned}&\mathbb{P}_B\left(n_{jk} < E[n_{jk}](1-\epsilon)\right) \leq \exp\left(-\epsilon^2(1-\delta)^2 n/2\right)\\ \Rightarrow&\mathbb{P}_B\left(\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\} \geq \frac{1}{p^{m+2}}\right)\\ &\leq \exp\left(-\epsilon^2(1-\delta)^2 n/2\right) , \end{aligned} \quad (8)$$

where we have substituted the expectation $E[n_{jk}]$. By considering probabilities over the missing values, we have

$$\begin{aligned}&\mathbb{P}_B\left(\sum_{j,k}\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\} > \frac{1}{p^m}\right)\\ &\leq \sum_{j,k}\mathbb{P}_B\left(\exp\left\{-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right\} > \frac{1}{p^{m+2}}\right)\\ &\leq p^2\exp\left(-\epsilon^2(1-\delta)^2 n/2\right)\\ &= \exp\left(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p)\right) , \end{aligned} \quad (9)$$

which completes the proof. ∎

3.1 Kendall's Tau with Missing Values

The following theorem shows that $\sup_{jk}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq O(\sqrt{\log p/n})$ with high probability.

Theorem 1 *For any $n \geq 1$, for any $m > 0$, and any $0 < \epsilon < 1$, with probability at least $(1-\frac{1}{p^m})(1-\exp(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p))$, we have*

$$\sup_{jk}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq \frac{\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}} . \quad (10)$$

Proof: Since $\hat{\tau}_{jk}$ is an unbiased estimator of τ_{jk}, $E[\hat{\tau}_{jk}] = \tau_{jk}$. Using (2), we have

$$\begin{aligned}&\mathbb{P}_X\left(\left|\hat{S}_{jk} - \Sigma^0_{jk}\right| > t\right)\\ &= \mathbb{P}_X\left(\left|\sin\left(\frac{\pi}{2}\hat{\tau}_{jk}\right) - \sin\left(\frac{\pi}{2}\tau_{jk}\right)\right| > t\right)\\ &\leq \mathbb{P}_X\left(|\hat{\tau}_{jk} - \tau_{jk}| > \frac{2}{\pi}t\right)\\ &\leq \exp\left(-\frac{n_{jk}t^2}{\pi^2}\right) , \end{aligned} \quad (11)$$

where the last inequality uses the Hoeffding bound for the U-statistics [12, 8]. Application of the union bound yields

$$\begin{aligned}&\mathbb{P}_X\left(\sup_{jk}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| > t\right)\\ &\leq \sum_{j,k}\exp\left(-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right) , \end{aligned} \quad (12)$$

where we have substituted $t = \frac{\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$. The bound in the above form is itself a random variable, and the elements of the sum are identically distributed but are not independent.

By considering probabilities over the missing values and using Lemma 3.1, we have

$$\begin{aligned}&\mathbb{P}_B\left(\mathbb{P}_X\left(\sup_{jk}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq t\right) \geq \left(1 - \frac{1}{p^m}\right)\right)\\ &\geq 1 - \exp\left(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p)\right) . \end{aligned} \quad (13)$$

Noting that the random variables (X, B) are independent completes the proof. ∎

3.2 Spearman's Rho with Missing Values

As we work on the n_{jk} effective samples wth values by disregarding missing values, we can leverage the analysis in [12] except n_{jk} is a random variable. Following [12], (3) can be rewritten as [7, 12]:

$$\begin{aligned}\hat{\rho}_{jk} &= \frac{3\sum_{i=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n}\text{sign}(x_i^j - x_s^j)(x_i^k - x_t^k)b_{ij}b_{ik}b_{sj}b_{sk}b_{tj}b_{tk}}{n_{jk}^3 - n_{jk}}\\ &= \frac{n_{jk}-2}{n_{jk}+1}U_{jk} + \frac{3}{n_{jk}+1}\hat{\tau}_{jk} . \end{aligned} \quad (14)$$

where $\hat{\tau}_{jk}$ is Kendall's tau statistics and U_{jk} is a 3rd-order U-statistics

$$U_{jk} = \frac{3\sum_{i\neq s\neq t}\text{sign}(x_i^j - x_s^j)(x_i^k - x_t^k)b_{ij}b_{ik}b_{sj}b_{sk}b_{tj}b_{tk}}{n_{jk}(n_{jk}-1)(n_{jk}-2)} . \quad (15)$$

Note $n_{jk} = \sum_{i=1}^{n} b_{ij}b_{ik}$ is a sum of n independent Bernoulli random variables $b_{ij}b_{ik}$ with $\mathbb{E}(n_{ij}) = (1-\delta)^2 n$.

Theorem 2 *For any $m > 0$, $0 < \epsilon < 1$, and*

$$n \geq \frac{36}{(m+2)(1-\epsilon)(1-\delta)^2 \log p}\,, \tag{16}$$

with probability at least $(1 - \frac{1}{p^m})(1 - \exp(-(\epsilon^2(1-\delta)^2 n/2 - 2\log p))$, we have

$$\sup_{jk} \left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}\,. \tag{17}$$

Proof: Let $0 < \alpha < 1$. According to (14), we have

$$\begin{aligned}\mathbb{P}_X(|\hat{\rho}_{jk} - \mathbb{E}(\hat{\rho}_{jk})| > t) &\leq \mathbb{P}_X(|U_{jk} - \mathbb{E}(U_{jk})| > \alpha t) \\ &+ \mathbb{P}_X\left(\frac{3}{n_{jk}+1}|\hat{\tau}_{jk} - \tau_{jk}| > (1-\alpha)t\right)\,.\end{aligned} \tag{18}$$

Since $-1 \leq \tau_{jk} \leq 1$, $|\hat{\tau}_{jk} - \tau_{jk}| \leq 2$, then

$$\begin{aligned}&\mathbb{P}_X\left(\frac{3}{n_{jk}+1}|\hat{\tau}_{jk} - \tau_{jk}| > (1-\alpha)t\right) \\ &\leq \mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t\right)\,.\end{aligned} \tag{19}$$

Applying Hoeffding's bound for U-statistics, we have

$$\begin{aligned}&\mathbb{P}_X(|U_{jk} - \mathbb{E}(U_{jk})| > \alpha t) \\ &\leq \exp\left(-2\left\lfloor\frac{n_{jk}}{3}\right\rfloor\frac{\alpha^2 t^2}{36}\right) = \exp\left(-\frac{n_{jk}\alpha^2 t^2}{54}\right)\,.\end{aligned} \tag{20}$$

Combining (19) and (20) yields

$$\begin{aligned}\mathbb{P}_X\left(|\hat{\rho}_{jk} - \mathbb{E}(\hat{\rho}_{jk})| > t\right) &\leq \exp\left(-\frac{n_{jk}\alpha^2 t^2}{54}\right) \\ &+ \mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t\right)\,.\end{aligned} \tag{21}$$

In particular, if $n_{jk} \geq \frac{6}{(1-\alpha)t}$, the second term on the RHS is 0. Since $\hat{\rho}_{jk}$ is a biased estimator, following [12], we use the following bias equation [26]:

$$\mathbb{E}\hat{\rho}_{jk} = \frac{6}{\pi(n_{jk}+1)}\left[\arcsin(\Sigma^0_{jk}) + (n_{jk}-2)\arcsin(\frac{\Sigma^0_{jk}}{2})\right]\,. \tag{22}$$

Note we only use n_{jk} effective number of samples. Thus,

$$\Sigma^0_{jk} = 2\sin\left(\frac{\pi}{2}\mathbb{E}\hat{\rho}_{jk} + a_{jk}\right)\,, \tag{23}$$

where

$$a_{jk} = \frac{\pi\mathbb{E}\hat{\rho}_{jk} - 2\arcsin(\Sigma^0_{jk})}{2(n_{jk}-2)}\,, |a_{jk}| \leq \frac{\pi}{n_{jk}-2}\,. \tag{24}$$

If $n_{jk} \geq \frac{6\pi}{t} + 2$, $|a_{jk}| \leq \frac{t}{6}$. Therefore, the analysis is simplified if $\inf_{jk} n_{jk} \geq c_0$ where

$$c_0 \geq \max\left\{\frac{6}{(1-\alpha)t}, \frac{6\pi}{t} + 2\right\}\,. \tag{25}$$

Setting $\alpha = \frac{3\sqrt{6}}{8}$, $t = \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$, we have

$$\begin{aligned}\frac{6}{(1-\alpha)t} &\leq \frac{24\pi}{t} = 6(1-\delta)\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}}\,, \\ \frac{6\pi}{t} + 2 &= \frac{3(1-\delta)}{2}\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}}\,.\end{aligned}$$

Therefore, we choose

$$c_0 = 6(1-\delta)\sqrt{\frac{1-\epsilon}{m+2}}\sqrt{\frac{n}{\log p}}\,. \tag{26}$$

Define an event $Z = \{\inf_{jk} n_{jk} \geq c_0\}$, and let $\bar{Z}$ be the complement of the event. Further, the event of interest is $Y = \left\{\sup_{j,k}\left|\hat{S}^{\tau}_{jk} - \Sigma^0_{jk}\right| \leq \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}\right\}$. Then, the probability of the event of interest can be lower bounded as:

$$\begin{aligned}P(Y) &= P(Y|Z)P(Z) + P(Y|\bar{Z})P(\bar{Z}) \\ &\geq P(Y|Z)P(Z)\,.\end{aligned} \tag{27}$$

Next, we focus on getting lower bounds to both $P(Z)$ and $P(Y|Z)$.

Note $n_{jk} = \sum_{i=1}^n b_{ij}b_{ik}$ and $\mathbb{E}[n_{jk}] = (1-\delta)^2 n$, using Chernoff bounds,

$$\mathbb{P}_B\left(n_{jk} < (1-\epsilon)(1-\delta)^2 n\right) \leq \exp\left(-\epsilon^2(1-\delta)^2 n/2\right)\,. \tag{28}$$

By the union bound,

$$\begin{aligned}&\mathbb{P}_B\left(\inf_{jk} n_{jk} < (1-\epsilon)(1-\delta)^2 n\right) \\ &\leq \exp\left(-\epsilon^2(1-\delta)^2 n/2 + 2\log p\right)\,,\end{aligned} \tag{29}$$

which is equivalent to

$$\begin{aligned}&\mathbb{P}_B\left(\inf_{jk} n_{jk} \geq (1-\epsilon)(1-\delta)^2 n\right) \\ &\geq 1 - \exp\left(-\epsilon^2(1-\delta)^2 n/2 + 2\log p\right)\,.\end{aligned} \tag{30}$$

If $(1-\epsilon)(1-\delta)^2 n \geq c_0$, i.e.,

$$n \geq \frac{36}{(m+2)(1-\epsilon)(1-\delta)^2 \log p}\,, \tag{31}$$

then

$$\mathbb{P}_B\left(\inf_{jk} n_{jk} \geq c_0\right) \geq 1 - \exp\left(-\epsilon^2(1-\delta)^2 n/2 + 2\log p\right)\,, \tag{32}$$

which gives a lower bound to $P(Z)$ as desired. Now, conditioned on Z, i.e., $\inf_{jk} n_{jk} \geq c_0$, we have $|a_{jk}| \leq \frac{t}{6}$, and $\mathbb{P}_X\left(\frac{6}{n_{jk}+1} > (1-\alpha)t \,\middle|\, Z\right) = 0$. Assuming n satisfies (31) and using (21), (23), we have

$$\begin{aligned}
&\mathbb{P}_X\left(|\hat{S}^{\rho}_{jk} - \Sigma^0_{jk}| > t \,\middle|\, Z\right) \\
&= \mathbb{P}_X\left(\left|2\sin\left(\frac{\pi}{6}\hat{\rho}_{jk}\right) - 2\sin(\frac{\pi}{6}\mathbb{E}\hat{\rho}_{jk} + a_{jk})\right| > t \,\middle|\, Z\right) \\
&\leq \mathbb{P}_X\left(\left|\frac{\pi}{3}\hat{\rho}_{jk} - \frac{\pi}{3}\mathbb{E}\hat{\rho}_{jk} - 2a_{jk}\right| > t \,\middle|\, Z\right) \\
&= \mathbb{P}_X\left(\left|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk} - \frac{6}{\pi}a_{jk}\right| > \frac{3t}{\pi} \,\middle|\, Z\right) \\
&\leq \mathbb{P}_X\left(|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk}| > \frac{3t}{\pi} - \left|\frac{6}{\pi}a_{jk}\right| \,\middle|\, Z\right) \\
&\leq \mathbb{P}_X\left(|\hat{\rho}_{jk} - \mathbb{E}\hat{\rho}_{jk}| > \frac{2t}{\pi} \,\middle|\, Z\right) \\
&\leq \exp\left(-\frac{2n_{jk}\alpha^2 t^2}{27\pi^2}\right) , \qquad (33)
\end{aligned}$$

where the conditioning on Z, i.e., $\{\inf_{j,k} n_{jk} \geq c_0\}$, has been dropped in the last inequality yielding an upper bound. Setting $\alpha = \frac{3\sqrt{6}}{8}, t = \frac{4\pi}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$, by the union bound, we have

$$\begin{aligned}
&\mathbb{P}_X\left(\sup_{jk}|\hat{S}^{\rho}_{jk} - \Sigma^0_{jk}| > t \,\middle|\, Z\right) \\
&\leq \sum_{j,k}\exp\left(-\frac{n_{jk}}{(1-\delta)^2(1-\epsilon)n}(m+2)\log p\right) , \qquad (34)
\end{aligned}$$

which is the same as (12). Using Lemma 3.1, we then have $P(Y|Z) \geq \left(1 - \frac{1}{p^m}\right)$. The result of the theorem then follows from (27) and (30). ∎

3.3 Plug-in CLIME Estimator

Since $\hat{S}$ ($\hat{S}^{\tau}$ or $\hat{S}^{\rho}$) satisfies (10) or (17) with high probability, choosing $\lambda_n \geq \frac{\pi\|\Omega^0\|_{L_1}}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$ or $\lambda_n \geq \frac{4\pi\|\Omega^0\|_{L_1}}{1-\delta}\sqrt{\frac{m+2}{1-\epsilon}}\sqrt{\frac{\log p}{n}}$ ensures that the conditions for consistency of the CLIME estimate $\hat{\Omega}$ are satisfied. The CLIME estimator considers the following family of precision matrices $\mathcal{U} = \mathcal{U}(M, q, s_0(p)) = \Big\{\Omega : \Omega \succ 0, \|\Omega\|_{L_1} \leq M, \max_{1\leq i\leq p}\sum_{j=1}^{p}|\omega_{ij}|^q \leq s_0(p)\Big\}$, for $0 \leq q < 1$. Then, the CLIME estimator has the following guarantees:

Theorem 3 *Let* $\Omega_0 \in \mathcal{U}(M, q, s_0(p))$. *If* $\lambda_n \geq \|\Omega_0\|_{L_1}\max_{ij}|\hat{\sigma}_{n,ij} - \sigma_{0,ij}|$, *then we have*

$$|\hat{\Omega}_n - \Omega_0|_{\infty} \leq 4\|\Omega_0\|_{L_1}\lambda_n , \qquad (35)$$

$$\|\hat{\Omega}_n - \Omega_0\|_2 \leq Cs_0(p)(4\|\Omega_0\|_{L_1})^{1-q}\lambda_n^{1-q} , \qquad (36)$$

$$\frac{1}{p}\|\hat{\Omega}_n - \Omega_0\|_F^2 \leq Cs_0(p)(4\|\Omega_0\|_{L_1})^{2-q}\lambda_n^{2-q} , \qquad (37)$$

where $C \leq 2(1 + 2^{1-q} + 3^{1-q})$ *is a constant.*

Note that deterministic bounds in Theorem 3 for precision estimation relies on $|\hat{\Sigma}_n - \Sigma_0|_{\infty} = \max_{i,j}|\hat{\sigma}_{n,ij} - \sigma_{0,ij}|$.

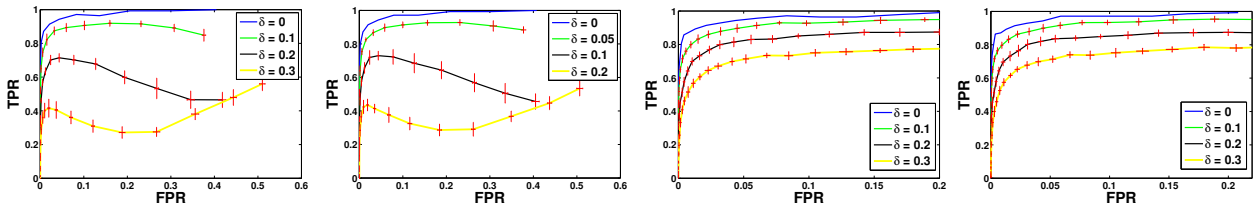
4 Experimental Results

We present experimental results of DoPinG on both synthetic datasets and real datasets to illustrate model performance. The first set of experiments on synthetic data illustrate the effect of sample size and percentage of missing data on model performance. Then we compare DoPinG with mGlasso on both synthetic data and climate dataset.

4.1 Synthetic Data

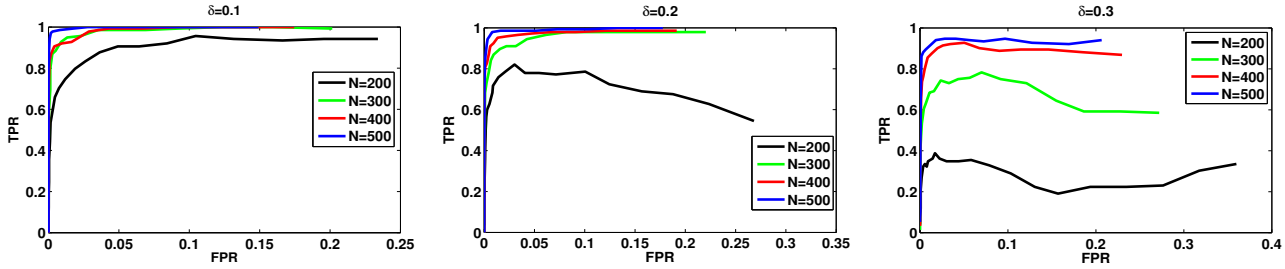
To generate synthetic data, we use the procedure described in [12]. First, a d-dimensional sparse graph $G = (V, E)$ is generated as follows: Let $V = \{1, ..., p\}$ correspond to variables $X = (X_1, ..., X_d)$. We associate each index j with a bivariate point $Y_j = (Y_j^{(1)}, Y_j^{(2)}) \in [0, 1]^2$ where each $Y_j^{(k)} \sim$ Unif$[0, 1]$, $k = 1, 2$, $j \in \{1, \cdots, d\}$. An edge is associated between vertices (i, j) with probability of $P((i, j) \in E) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|y_i - y_j\|^2}{0.25}\right)$ where $y_j = (y_j^{(1)}, y_j^{(2)})$ is the observation of Y_j and $\|\ .\ \|$ denotes the Euclidean distance. The maximum degree of the graph is limited to 4. Thereafter, n samples are drawn from $NPN_d(f^0, \Sigma^0)$ where f^0 is the Gaussian CDF Transformation with mean 0.05 and standard deviation 0.4. Here, we choose $n = 200$, $p = 100$, and $\delta \in \{0.1, 0.2, 0.3\}$. The final results shown below are averages over 10 experimental runs for both Kendall's tau and Spearman's rho. The ROC curve is generated by varying the tuning parameter λ in the CLIME and calculating the corresponding False Positive Rate (FPR) and True Positive Rate (TPR) [12].

First, we directly run Algorithm 1 using $\hat{S}$ ($\hat{S}^{\tau}$ or $\hat{S}^{\rho}$) estimated using Kendall's tau and Spearman's rho. The ROC curve with different probabilities of missing values is plotted in Figure 1. We observe that the performance of Kendall's tau and Spearman's rho is almost the same for the same percentage of missing values. Note that the tuning parameter λ controls the sparsity of the estimated graph, i.e., a small value of λ provides a dense graph. When λ is large enough the predicted edges are all among the correct edges leading to a zero FPR. By decreasing λ, false edges that are not in the original graph are added, i.e., increasing FPR and saturating TPR. It shows that the estimator is conservative



(a) Kendall no projection (b) Spearman no projection (c) Kendall, projection (d) Spearman, projection

Figure 1: (a,b) ROC curves without projection ($\hat{S}$ need not be positive semi-definite), (c,d) ROC curves with projection ($\hat{S}$ is positive semi-definite) with $n = 200$ and under different missing probabilities ($\delta = 0.1 - 0.3$). By increasing number of observed data (smaller δ), the ROC curve approaches the ROC curve of no-missing data ($\delta = 0$).



(a) $\delta = 10\%$ (b) $\delta = 20\%$ (c) $\delta = 30\%$

Figure 2: ROC curve with $\delta = 0.1, 0.2, 0.3$, $p = 100$, and different number of samples (n). For a fixed value of δ, with increasing number of samples, the higher TP rates is obtained.

in adding edges. Figure 1 also illustrates that increasing number of missing values (increasing δ) deteriorates model performance, while increasing variance of estimate.

As mentioned in section 2.3, the estimated correlation matrix $\hat{S}$ may be not positive semi-definite. Therefore, we project $\hat{S}$ into the positive semi-definite (PSD) cone, and execute Algorithm 1 using the PSD matrix. Figures 1 (c,d) plot the ROC curve with projection for Kendall's tau and Spearman's rho respectively. For small δ, e.g. $\delta = 0.1$, to some degree, the performances with and without projection are similar. However, when more values are missing, PSD projection greatly improves performance. Increasing percentage of missing values lead to more and larger negative eigenvalues in $\hat{S}$, and performance worsens for higher δ. Note that our analysis shows that the *effective* sample size is $(1-\delta)^2 n$, and decrease of the recovery rate (TPR) with decreasing *effective* sample size is in accordance with our analysis. In other words, for a fixed n the *effective* sample size is smaller for a larger value of δ and therefore, DoPinG has a worse performance with larger value of δ.

Figure 2 shows the effect of sample size n with different value of δ on the performance without projection. Under higher percentage of missing values (Figure 2(c)), the performance of the method suffers much more with low sample size, compared to data with lower percentage of missing entries (Figure 2(a)). In particular, with a sample size $n = 200$ and 30% of missing data, the *effective* sample size is ~ 100 while with 10% of missing data, the *effective* sample size is ~ 160. As a result, to achieve similar recovery rates (TPR,FPR), higher sample size is needed when more

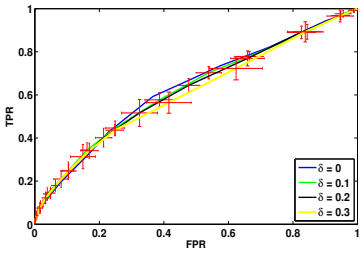


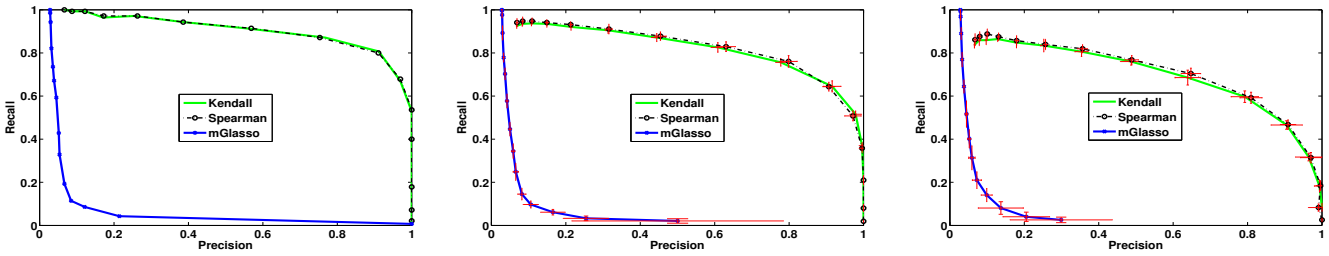
Figure 3: ROC curve of mGlasso with $n = 200$ and different missing probabilities. mGlasso has a worse performance on non-Gaussian data compared to DoPinG (Figure 1).

percentage of the data is missing.

We compare DoPinG with mGlasso [9] on the synthetic data. The ROC curve of mGlasso is plotted in Figure 3. Since mGlasso is designed primarily for Gaussian data, Figure 3 clearly illustrates that mGlasso is not suitable for non-Gaussian data. We also plot the precision and recall curve with different probabilities of missing values ($\delta = 0, 0.1, 0.2$) in Figure 4. The performance of DoPinG is significantly better than mGlasso.

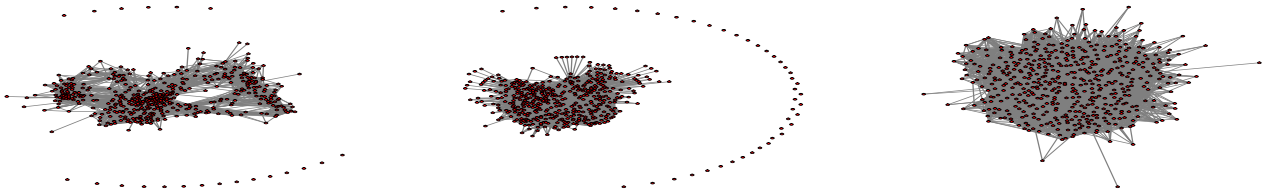
4.2 Climate Data

We compare DoPinG (Spearman's rho) and mGlasso on Climate data. The climate dataset that we use is obtained from the CMIP5 archive, where we use the temperature predicted over land locations by a climate model. We reduce the resolution of the data, since we use it only for illustrative purposes, so that the data contains 500 locations (di-



(a) $\delta = 0$ (b) $\delta = 10\%$ (c) $\delta = 20\%$

Figure 4: Precision and Recall Curve with different δ. DoPinG is significantly better than mGlasso for non-Gaussian data.



(a) DoPinG (12240 edges) (b) mGlasso (8778 edges) (c) mGlasso (11860 edges)

Figure 5: The graph discovered by DoPinG and mGlasso.

Table 1: Edges dicovered by DoPinG and mGlasso on Climate Data. $>$ denotes the number of edges in DoPinG graph but not in mGlasso graph. $<$ is on the contrary.

Edge No.		Edge Diff	
DoPinG	mGlasso	$>$	$<$
12240	8778	7942	4480
12240	11860	7534	7154

mensionality), and yearly averaged samples over 100 years (sample size =100). We randomly remove $\delta = 20\%$ of the entries. We try different λ and report the results which have similar number of edges. In particular, we pick the graph with 12740 edges for DoPinG ($\lambda = 0.02$) as illustrated in Figure 5(a). We pick two graphs for mGlasso. One has 8778 edges ($\lambda = 0.001$) and the other has 11860 edges ($\lambda = 0.002$), as shown in Figure 5(b) and 5(c) respectively. It seems that DoPinG discovers some interesting sparsity patterns while mGlasso graphs are messy. In Table 1, we present the difference between DoPinG graph and mGlasso graph. With similar total number of edges, DoPinG graph shows more structure than mGlasso graph. We plan to further investigate this behavior in future work.

5 Conclusions

In this paper, we propose double plugin Gaussian (DoPinG) copula estimators to deal with non-Gaussian data with missing values. DoPinG estimates the sparse precision matrix corresponding to *non-paranormal* distributions by directly estimating nonparametric correlations, including Kendall's tau and Spearman's rho. DoPinG uses two plugin procedures, leveraging existing sparse precision estimators. DoPinG consists of three steps: (1) estimate nonparametric correlations by disregarding missing values; (2) estimate the non-paranormal correlation matrix directly based on nonparametric correlations like Kendall's tau and Spearman's rho; (3) plug the estimated correlation matrix into existing sparse precision estimators to yield the sparse precision matrix. We prove that DoPinG copula estimators consistently estimate the non-paranormal correlation matrix at a rate of $O(\frac{1}{(1-\delta)}\sqrt{\frac{\log p}{n}})$, where δ is the probability of missing values. Through experiments we illustrate that by increasing number of missing values (increasing δ), the performance of the method get worse and the standard deviation is increasing in consistent with the theory. The performance of Kendall's tau and Spearman's rho is almost the same for the same percentage of missing values. Experimental results on non-Gaussian data show that DoPinG is significantly better than estimators like mGlasso, which are primarily designed for Gaussian data.

Acknowledgment

The research was supported by NSF grants IIS- 0916750, IIS-0953274, IIS-1029711, CNS-1314560, by NASA grant NNX12AQ39A. A.B. acknowledges support from IBM and Yahoo. H.W. acknowledges the support of DDF (2013-2014) from the University of Minnesota.

References

[1] O. Banerjee, L. E. Ghaoui, and A. dAspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 9:2261–2286, 2008.

[2] S. Boyd, E. Chu N. Parikh, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation and Trends Machine Learning*, 3(1), 2011.

[3] T. Cai, W. Liu, and H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Preprint*, 2012.

[4] T. Cai, C.H. Zhang, and H. Zhou. A constrained l_1 minimization approach to sparse precision matrix estimation. *American Statistical Association*, 106:594–607, 2011.

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

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

[7] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19:293–325, 1948.

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

[9] M. Kolar and E. Xing. Estimating sparse precision matrices from data with missing values. In *ICML*, 2012.

[10] W. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[11] R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1987.

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

[13] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *JMLR*, 10:2295–2328, 2009.

[14] H. Liu and L. Wang. Tiger: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Preprint*, 2012.

[15] P. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *NIPS*, 2012.

[16] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *ArXiv*, 2012.

[17] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

[18] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[19] N. Stadler and P. Buhlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, pages 1–17, 2009.

[20] H. Tsukahara. Efficient estimation in the bivariate normal copula model: Normal margins are least-favorable. *Bernoulli*, 3:55–77, 1997.

[21] H. Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33:357–375, 2005.

[22] H. Wang and A. Banerjee. Online alternating direction method. In *ICML*, 2012.

[23] H. Wang, A. Banerjee, C. Hsieh, P. Ravikumar, and I. Dhillon. Large scale distributed sparse precision estimation. In *NIPS*, 2013.

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

[25] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *JMLR*, 11, 2010.

[26] D. Zimmerman, B. Zumbo, and R. Williams. Bias in estimation and hypothesis testing of correlation. *Transformation*, 24:133–158, 2003.