Robust Estimation of Gaussian Copula Causal Structure from Mixed Data with Missing Values

Ruifei Cui, Perry Groot, Tom Heskes
Institute for Computing and Information Sciences
Radboud University
Nijmegen, Netherlands
Email: {r.cui, perry.groot, t.heskes}@science.ru.nl

***Abstract*—We consider the problem of causal structure learning from data with missing values, assumed to be drawn from a Gaussian copula model. First, we extend the 'Rank PC' algorithm, designed for Gaussian copula models with purely continuous data (so-called *nonparanormal* models), to incomplete data by applying rank correlation to pairwise complete observations and replacing the sample size with an effective sample size in the conditional independence tests to account for the information loss from missing values. The resulting approach works when the data are *missing completely at random* (MCAR). Then, we propose a Gibbs sampling procedure to draw correlation matrix samples from mixed data under *missingness at random* (MAR). These samples are translated into an average correlation matrix, and an effective sample size, resulting in the 'Copula PC' algorithm for incomplete data. Simulation study shows that: 1) the usage of the effective sample size significantly improves the performance of 'Rank PC' and 'Copula PC'; 2) 'Copula PC' estimates a more accurate correlation matrix and causal structure than 'Rank PC' under MCAR and, even more so, under MAR. Also, we illustrate our methods on a real-world data set about gene expression.**

I. Introduction

Causal structure learning [1], or causal discovery, aims to learn underlying directed acyclic graphs (DAG), in which the vertices denote random variables and the edges represent causal relations among the variables. It is a useful tool for multivariate analysis and has been widely studied in the past decade [2]–[5]. Constraint-based methods, e.g., the PC algorithm and the FCI algorithm [2], have attracted extensive attention and generated many recent improvements [3], [4], [6], [7], yielding better search strategies and interpretability. Since all these algorithms share the adjacency search of the PC algorithm as a common first step, any improvements to PC can be directly transferred to the others. Therefore, we focus our analysis on the PC algorithm in this paper.

The adjacency search of the PC algorithm starts with a fully connected undirected graph, and then recursively removes the edges according to conditional independence decisions. For testing the conditional independence, the PC algorithm requires the correlation matrix and the sample size as input. The sample size is necessary: the higher the sample size, the more reliable the estimated correlation matrix, and the more easily the null hypothesis of conditional independence gets rejected (see Equation (1)). When applied to Gaussian data, the standard PC algorithm estimates the correlation matrix based on Pearson correlations. Harris & Drton [4] extended the PC algorithm to *nonparanormal* models, i.e., Gaussian copula models with purely continuous marginal distributions, by replacing the Pearson correlations with rank-based correlations. Cui et al. [7] further extended the PC algorithm to mixed discrete and continuous data assumed to be drawn from a Gaussian copula model. However, all these approaches were based on the assumption that the data are fully observed.

In practice, all branches of experimental science are plagued by data with missing values [8], [9], e.g., failure of sensors or drop-outs of subjects in a longitudinal study. In this paper, we target to generalize the PC algorithm to settings where the data are still assumed to be drawn from a Gaussian copula model, but with some missing values. For this, we need to estimate the underlying correlation matrix and the 'effective sample size' from incomplete data. The notion 'effective sample size', typically smaller than or equal to the sample size, was proposed in [7] to account for the information loss incurred by discrete variables. In this paper, we use it to account for the information loss incurred by missing values, acting as if the estimated correlations on incomplete data are in fact estimated from a smaller size of equivalent complete data.

In nonparanormal cases, Wang et al. [10] proposed to apply rank correlation to pairwise complete observations for estimating the correlation matrix, which is then plugged into existing procedures for inferring the underlying undirected graphical structure. In this paper, we transfer this idea to causal structure learning, where this estimator is used for the correlation matrix and the number of pairwise complete observations is taken as the effective sample size. This extends the 'Rank PC' algorithm to incomplete data. However, this resulting approach only applies to nonparanormal data under *missingness completely at random* (MCAR), which is a pretty strong assumption [11]. By contrast, we prefer an approach that is valid for both nonparanormal and mixed data under a less restrictive assumption, *missingness at random* (MAR) [11], [12]. To this end, we propose a Gibbs sampling procedure to draw correlation matrix samples from the posterior distribution given mixed continuous and discrete data with missing values. Then, following the idea of the 'Copula PC' algorithm [7], these Gibbs samples are translated into an average correlation matrix and an effective sample size, which are input to the standard PC algorithm for causal discovery. The difference is that now the effective sample size accounts for information

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.101

loss incurred by both missing values and discrete variables.

The rest of this paper is organized as follows. Section II reviews some necessary background knowledge. Section III and Section IV introduce the 'Rank PC' algorithm and the 'Copula PC' algorithm for incomplete data respectively. Section V evaluates the proposed algorithms on simulated data, and provides an illustration on real-world data set about gene expression. Section VI concludes this paper.

II. Preliminaries

A. Missingness Mechanism

Following [11], let $\boldsymbol{Y} = (y_{ij}) \in \mathbb{R}^{n\times p}$ be a data matrix with the rows representing independent samples, and $\boldsymbol{R} = (r_{ij}) \in \{0,1\}^{n\times p}$ be a matrix of indicators, where $r_{ij} = 1$ if y_{ij} was observed and $r_{ij} = 0$ otherwise. $\boldsymbol{Y}$ consists of two parts, $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{miss}$, where $\boldsymbol{Y}_{obs}$ contains the observed elements in $\boldsymbol{Y}$ and $\boldsymbol{Y}_{miss}$ the missing elements. When the missingness does not depend on the observed values, i.e., $P(\boldsymbol{R}|\boldsymbol{Y},\theta) = P(\boldsymbol{R}|\theta)$ with θ denoting unknown parameters, the data are said to be *missing completely at random* (MCAR), which is a special case of a more realistic assumption called *missing at random* (MAR). MAR allows the dependency between missingness and observed values, i.e., $P(\boldsymbol{R}|\boldsymbol{Y},\theta) = P(\boldsymbol{R}|\boldsymbol{Y}_{obs},\theta)$. For example, all people in a group are required to take a blood pressure test at time point 1, while only those whose values at time point 1 lie in the abnormal range need to take the test at time point 2. This results in some missing values at time point 2 that are MAR.

B. Gaussian Copula Model

Definition 1 (Gaussian Copula Model)**.** Consider a latent random vector $\boldsymbol{Z} = (Z_1,\ldots,Z_p)^T$ and an observed random vector $\boldsymbol{Y} = (Y_1,\ldots,Y_p)^T$, satisfying the conditions

$$\boldsymbol{Z} \sim \mathcal{N}(0, C),$$
$$Y_j = F_j^{-1}\big[\Phi(Z_j)\big], \forall j = 1,\ldots,p,$$

where C denotes the correlation matrix of $\boldsymbol{Z}$, $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian, and ${F_j}^{-1}(t) = \inf\{y : F_j(y) \geq t\}$ is the pseudo-inverse of a cumulative distribution function F_j. Then this model is called a *Gaussian copula model* with correlation matrix C and univariate margins F_j.

This model provides an elegant way to analyze diverse types of variables, say binary, ordinal, and continuous [13]. Also, the associations between observed variables in this model are parameterized separately from their marginal distributions, which supplies much flexibility in multivariate analysis [13].

C. Causal Discovery

A graphical model is a graph $G = (V, E)$, where the vertices $(X_i : X_i \in V)$ denote random variables and the edges E represent dependence structure among the variables. A graph is *directed* if it just contains directed edges and *undirected* if all edges are undirected. A graph that contains both directed and undirected edges is called a *partially directed* graph. Graphs without directed cycles (e.g., $X_i \to X_j \to X_i$) are *acyclic*. We refer to a graph as a *Directed Acyclic Graph* (DAG) if it is both directed and acyclic. If there is a directed edge $X_i \to X_j$, we say that X_i is a parent of X_j .

A probability distribution over a random vector $\boldsymbol{X}$ with $X_i \in V$ is said to be Markov w.r.t. a DAG $G = (V, E)$, if $\boldsymbol{X}$ satisfies the *Causal Markov Condition*: each variable in G is independent of its non-descendants given its parents, which is also implied by so-called *d-separation* [1]. A distribution is *faithful* w.r.t. a DAG if there are no conditional independencies in the distribution that are not encoded via d-separation.

Several DAGs may, via d-separation, correspond to the same set of conditional independencies. The set of such DAGs is called a *Markov Equivalence Class*, which can be represented by a *completed partially directed acyclic graph* (CPDAG) [14]. Arcs in a CPDAG imply a cause-effect relationship between pairs of variables since the same arc appears in all members of the CPDAG. An undirected edge $X_i - X_j$ in a CPDAG indicates that some of its members contain an arc $X_i \to X_j$ while others contain an arc $X_j \to X_i$. Causal discovery aims to learn the Markov equivalence class of the DAG G from observations of $\boldsymbol{X}$.

D. PC Algorithm

The PC algorithm [2], a reference algorithm for causal discovery, consists of two stages: adjacency search and orientation. Starting with a fully connected undirected graph, the adjacency search recursively removes the edges according to conditional independence decisions, yielding the skeleton and separation sets. The orientation first directs the unshielded triples according to the separation sets, and then directs as many of the remaining undirected edges as possible by applying the orientation rules repeatedly.

A key part of the procedure is to test conditional independence. When a random vector $\boldsymbol{X} \sim \mathcal{N}(0, C)$, the PC algorithm considers the so-called partial correlation, denoted by $\rho_{uv|S}$, which can be estimated through the correlation matrix C [15]. Specifically, given observations of $\boldsymbol{X}$ and significance level α, classical decision theory yields

$$X_u \perp\!\!\!\perp X_v|\boldsymbol{X}_S \Leftrightarrow \tag{1}$$
$$\sqrt{n - |S| - 3}\left|\frac{1}{2}\log\left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}}\right)\right| \leq \Phi^{-1}(1-\alpha/2),$$

where $u \neq v$, $S \subseteq \{1,\ldots,p\}\backslash\{u, v\}$. Hence, the PC algorithm requires the sample correlation matrix $\hat{C}$ (to estimate $\rho_{uv|S}$) and the sample size n as input.

III. Rank PC Algorithm for Incomplete Data

Our procedure of the 'Rank PC' algorithm for data with missing values consists of three steps: 1) estimate rank correlations based on pairwise complete observations; 2) estimate the underlying correlation matrix and the effective sample size; 3) plug these into the standard PC algorithm for causal discovery.

Since the two typical rank correlations, Kendall's τ and Spearman's ρ, are similar in our analysis, we focus our attention on Kendall's τ in this paper. Given the data matrix $\boldsymbol{Y}$

and indicator matrix $\boldsymbol{R}$, we compute the Kendall's τ between Y_j and Y_k on samples which have observed values for both the two variables, i.e.,

$$\hat{\tau}_{jk} = \frac{2}{\hat{n}_{jk}(\hat{n}_{jk}-1)} \sum_{1\le i<i'\le n} r_{ij}r_{ik}r_{i'j}r_{i'k}K(y_i, y_{i'}), \quad (2)$$

where $K(y_i, y_{i'}) = \operatorname{sign}\left((y_{ij}-y_{i'j})(y_{ik}-y_{i'k})\right)$ and $\hat{n}_{jk} = \sum_{i=1}^{n} r_{ij}r_{ik}$, which is the number of pairwise complete observations for variables Y_j and Y_k.

Proposition 1 (refer to [16], [17])**.** *Assuming $\boldsymbol{X}$ follows a nonparanormal distribution with correlation matrix C, we have $C_{jk} = \sin\left(\frac{\pi}{2}\tau_{jk}\right)$.*

Motivated by Proposition 1, we consider the estimator $\hat{S}^{\tau} = (\hat{S}^{\tau}_{jk})$:

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

When translating the number of pairwise complete observations $\hat{n}_{jk}$ (see Equation (2)) into an effective sample size to be used in the conditional independence tests of the PC algorithm, we compare two schemes.

Scheme 1: We take the average over all the $\hat{n}_{jk}$'s, i.e.,

$$\hat{n} = \frac{2}{p(p-1)} \sum_{1\le j<k\le p} \hat{n}_{jk}\,.$$

We refer to this estimator $\hat{n}$ as the global effective sample size. In this scheme, all the conditional independence tests share the same effective sample size.

Scheme 2: We give a different effective sample size to different tests, since each test relies on a local structure involving only part of the variables. In this case, we rewrite the conditional independence testing criteria to

$$X_u \perp\!\!\!\perp X_v | \boldsymbol{X}_S \Leftrightarrow \quad (3)$$
$$\sqrt{\hat{n}_{uv|S} - |S| - 3}\left|\frac{1}{2}\log\left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}}\right)\right| \le \Phi^{-1}(1-\alpha/2),$$

where $\hat{n}_{uv|S}$ is defined as

$$\hat{n}_{uv|S} = \frac{2}{q(q-1)} \sum_{\substack{j,k\in\{u,v,S\} \\ j<k}} \hat{n}_{jk},$$

with $q = 2 + |S|$. We refer to $\hat{n}_{uv|S}$ as the local effective sample size.

In the last step, we take the estimated correlation matrix $\hat{S}^{\tau}$ and the global (or local) effective sample size as input to the standard PC algorithm for causal discovery.

The convergence rate of this rank-based correlation estimator $\hat{S}^{\tau}$ has been derived in the presence of missing values (Theorem 1 in [10]). Building upon this convergence property, one could also derive the error bound of 'Rank PC' for incomplete data following the same line of reasoning as in [4].

IV. Copula PC Algorithm for Incomplete Data

In this section, we extend the 'Copula PC' algorithm to incomplete data. It includes three steps: 1) apply a Gibbs sampler to draw correlation matrix samples from the posterior distribution given data with missing values (Section IV-A); 2) use these samples to estimate the underlying correlation matrix and the effective sample size (Section IV-B); 3) plug the estimated correlation matrix and effective sample size into the standard PC algorithm for causal discovery.

A. Gibbs Sampling for Data with Missing Values

We choose Σ from an inverse-Wishart distribution, denoted by $\mathcal{W}^{-1}(\Sigma; \Psi, \nu)$, and write $P(C) = \mathcal{PW}^{-1}(C; \Psi, \nu)$, where $C = (C_{jk})$ with $C_{jk} = \Sigma_{jk}/\sqrt{\Sigma_{jj}\Sigma_{kk}}$. Then this distribution on correlation matrix C is called a *projected inverse-Wishart distribution*, which is the conjugate prior for Gaussian models. Specifically, when we choose the prior $P(C) = \mathcal{PW}^{-1}(C; \Psi_0, \nu_0)$, the posterior given data $\boldsymbol{Z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_n})^T$ reads

$$P(C|\boldsymbol{Z}) = \mathcal{PW}^{-1}(C; \Psi_0 + \boldsymbol{Z}^T\boldsymbol{Z}, \nu_0 + n)\,. \quad (4)$$

For Gaussian copula models with missing values, we cannot observe the random vector $\boldsymbol{Z}$ directly (refer to Definition 1), but an idea is to first obtain the Gaussian pseudo-data from the observed data (i.e., $\boldsymbol{Y}$) and then do inference for C. We use a Gibbs sampling procedure to implement this idea.

Let $\boldsymbol{Z} = (z_{ij}) \in \mathbb{R}^{n\times p}$ be the Gaussian pseudo-data implied by $\boldsymbol{Y}$, thus $\boldsymbol{Z}$ has two parts as well, $\boldsymbol{Z}_{obs}$ and $\boldsymbol{Z}_{miss}$. As initialization of our Gibbs sampling procedure, we propose to obtain the Gaussian pseudo-data of non-missing values $\boldsymbol{Z}_{obs}$. For this, we substitute the empirical cumulative distribution function based on non-missing data $\boldsymbol{Y}_{obs}$:

$$z_{ij} = \Phi^{-1}\left[\frac{\sum_{d=1}^{n} r_{dj}\mathbb{1}(y_{dj} < y_{ij})}{\sum_{d=1}^{n} r_{dj} + 1}\right], \text{ if } r_{ij} = 1\,, \quad (5)$$

where $\mathbb{1}(\cdot)$ is the indicator function.

For nonparanormal data with missing values completely at random, each marginal distribution of $\boldsymbol{Z}_{obs}$ can approximately represent the underlying true distribution. Then we iterate the following two steps to impute missing values (step 1, a conditional Gaussian) and draw correlation matrix samples from the posterior (step 2, a projected inverse-Wishart) [18]:

1) $\boldsymbol{Z}_{miss} \sim P(\boldsymbol{Z}_{miss}|\boldsymbol{Z}_{obs}, C)$;
2) $C \sim P(C|\boldsymbol{Z}_{obs}, \boldsymbol{Z}_{miss})$.

However, for mixed data under MAR, the initialization shown in Equation (5) is no longer sufficient for two reasons: 1) tied observations may occur, making the ranks no longer well-defined; 2) the missing values in one variable may depend on the values of others. These differentiate the obtained marginal distributions from the underlying true distributions. Hence, we need a different strategy to obtain $\boldsymbol{Z}_{obs}$.

For this, we borrow the idea of the so-called extended rank likelihood [13], derived as follows. Since the transformation $Y_j = F_j^{-1}[\Phi(Z_j)]$ is non-decreasing, observing

Algorithm 1 Gibbs sampler for mixed data under MAR

1: **Step 1**: $\boldsymbol{Z}_{obs} \sim P(\boldsymbol{Z}_{obs}|\boldsymbol{Z}_{obs} \in D(\boldsymbol{Y}_{obs}), C)$.
2: **for** $j \in \{1, \ldots, p\}$ **do**
3: $\quad \boldsymbol{v}^T = C_{[j,-j]} C^{-1}_{[-j,-j]}$
4: $\quad \sigma_j^2 = C_{[j,j]} - \boldsymbol{v}^T C_{[-j,j]}$
5: $\quad$ **for** $y \in \text{unique}\{y_{1,j}, \ldots, y_{n,j}\}$ **do**
6: $\quad\quad z_l = \max\{z_{i,j} : y_{i,j} < y\}$
7: $\quad\quad z_u = \min\{z_{i,j} : y < y_{i,j}\}$
8: $\quad\quad$ **for** i such that $y_{i,j} = y$ **do**
9: $\quad\quad\quad \mu_{i,j} = \boldsymbol{Z}_{[i,-j]} \times \boldsymbol{v}$
10: $\quad\quad\quad$ Draw $u_{i,j}$ from $\mathcal{U}\big(\Phi\big[\frac{z_l - \mu_{i,j}}{\sigma_j}\big], \Phi\big[\frac{z_u - \mu_{i,j}}{\sigma_j}\big]\big)$
11: $\quad\quad\quad z_{i,j} = \mu_{i,j} + \sigma_j \times \Phi^{-1}(u_{i,j})$
12: $\quad\quad$ **end for**
13: $\quad$ **end for**
14: **end for**
15: **Step 2**: $\boldsymbol{Z}_{miss} \sim P(\boldsymbol{Z}_{miss}|\boldsymbol{Z}_{obs}, C)$.
16: $\boldsymbol{Z} = (\boldsymbol{Z}^T - \boldsymbol{\mu})^T$, with $\boldsymbol{\mu}$ the mean vector of $\boldsymbol{Z}$.
17: **Step 3**: $C \sim P(C|\boldsymbol{Z}_{miss}, \boldsymbol{Z}_{obs})$.

$\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n,j})^T$ implies a partial ordering on $\boldsymbol{z}_j = (z_{1,j}, \ldots, z_{n,j})^T$, i.e., $\boldsymbol{z}_j$ must lie in

$$D(\boldsymbol{y}_j) = \{\boldsymbol{z}_j \in \mathbb{R}^n : y_{i,j} < y_{k,j} \Rightarrow z_{i,j} < z_{k,j}\} .$$

Therefore, observing $\boldsymbol{Y}$ suggests that $\boldsymbol{Z}$ must be in

$$D(\boldsymbol{Y}) = \{\boldsymbol{Z} \in \mathbb{R}^{n \times p} : \boldsymbol{z}_j \in D(\boldsymbol{y}_j), \forall j = 1, \ldots, p\} .$$

Taking the occurrence of this event as the data, one can compute the following likelihood

$$\begin{aligned} P(\boldsymbol{Z} \in D(\boldsymbol{Y})|C, F_1, \ldots, F_p) &= \int_{D(\boldsymbol{Y})} p(\boldsymbol{Z}|C) d\boldsymbol{Z} \\ &= P(\boldsymbol{Z} \in D(\boldsymbol{Y})|C), \end{aligned}$$

which is independent of the margins F_j. Then inference for C proceeds by iterating the following two steps:

1) $\boldsymbol{Z} \sim P(\boldsymbol{Z}|\boldsymbol{Z} \in D(\boldsymbol{Y}), C)$;
2) $C \sim P(C|\boldsymbol{Z})$.

The strong posterior consistency for C under the extended rank likelihood has been proven in [19]. We now use this method to obtain $\boldsymbol{Z}_{obs}$ from $\boldsymbol{Y}_{obs}$, resulting in the Gibbs sampler in Algorithm 1. Note that line 16 in Algorithm 1 needs to relocate the data such that the mean of each coordinate of $\boldsymbol{Z}$ is zero. This is necessary for the algorithm to be sound because the mean may shift when missing values depend on the observed data (MAR). This Gibbs sampler can be implemented using the R package **sbgcop** [20], where line 16 in Algorithm 1 should be added to guarantee that the procedure is consistent also under MAR [1].

B. Estimating the Underlying Correlation Matrix and the Effective Sample Size

By iterating the steps in Algorithm 1, we can draw samples of the correlation matrix, denoted by $\{C^{(1)}, \ldots, C^{(m)}\}$. The mean over all the samples is a natural estimate of the underlying correlation matrix $\hat{C}$, i.e.,

$$\hat{C} = \frac{1}{m} \sum_{i=1}^{m} C^{(i)} .$$

We refer to the estimator as the copula estimator for the correlation matrix.

For estimating the effective sample size, while it is straightforward for the pairwise deletion method (the one we used in Section III), a different strategy in the current case is needed.

The projected inverse-Wishart distribution has a property that is summarized in Theorem 1 (see [7] for the proof), showing the relationship between the mean, variance and degrees of freedom.

Theorem 1. *Consider a p-dimensional random matrix C. If $P(C) = \mathcal{PW}^{-1}(C; \Psi, \nu)$, we have*

$$\text{Var}\left[C_{jk}\right] \approx \frac{(1 - (\text{E}\left[C_{jk}\right])^2)^2}{\nu} ,$$

for each off-diagonal element C_{jk} and large ν ($\gg p$).

In Equation (4), since generally $\nu_0 \ll n$, the posterior degrees of freedom $\nu_0 + n \approx n$. From Theorem 1, following the idea in [7], the effective sample size for the correlation C_{jk} (denoted by $\hat{n}_{jk}$ for clarity since it can vary for different combinations of j and k) reads

$$\hat{n}_{jk} = \frac{(1 - (\text{E}_n[C_{jk}])^2)^2}{\text{Var}_n[C_{jk}]}, \forall j \neq k ,$$

where $\text{E}_n[C_{jk}]$ and $\text{Var}_n[C_{jk}]$ denote respectively the mean and variance estimated through the correlation matrix samples drawn from an n-size mixed data set with missing values.

Compared to fully observed and continuous data, the difference now is that the correlation matrix samples contain some additional variance incurred by missing values and ties in discrete variables, which results in a reduced degrees of freedom used as our effective sample size.

When applying the effective sample size to conditional independence tests, we also compare the two schemes as shown in Section III.

V. Experiments

A. Simulation Setup

Following [21], we generate random DAGs and simulate normally distributed samples that are faithful to them. This procedure is implemented via functions 'randomDAG' and 'rmvDAG' in the R package **pcalg** [22].

Missing values with a certain percentage δ_j in a variable (the jth variable) are created following the procedure in [23]:

- Under MCAR, $\forall i, j$, $z_{i,j}$ is missing if $r_{i,j} = 0$ where $r_{i,j} \sim \text{Bern}(1 - \delta_j)$.
- Under MAR, for $j = 1, \ldots, \lfloor p/2 \rfloor$, $i = 1, \ldots, n$: $z_{i,2*j}$ is missing if $z_{i,2*j-1} < \Phi^{-1}(\delta_j)$.

We randomly draw δ_j from a uniform distribution in the interval $[0, 0.5]$, i.e., $\delta_j \sim \mathcal{U}(0, 0.5), \forall j$.

[1]The code is also provided in https://www.dropbox.com/sh/s8cwbavt9kx14ga/AAB6Y7iNwQjwNinM8dcmox4qa?dl=0

We restrict the number of variables to $p = 20$. The significance level in the standard PC algorithm is set to $\alpha = 0.01$ and the sparseness parameter in generating DAGs to $s = 2/(p-1)$, such that the average neighbors of each node is two [21]. For the Gibbs sampling, we abandon the first 100 samples (burn-in) and save the next 100.

B. Causal Discovery on Simulations

In this subsection, we assess the justification of the usage of the effective sample size in causal discovery, and compare the 'Rank PC' with the 'Copula PC' algorithm on simulated data. The true positive rate (TPR) and the false positive rate (FPR) are used to evaluate the estimated skeleton, while the structural Hamming distance (SHD) [24] evaluates the CPDAG. A higher TPR, a lower FPR, and a smaller SHD imply better performance.

The results for $n = 1000$ on nonparanormal data are shown in Figure 1, where the combinations of CoPC with SS, GESS, and LESS denote the Copula PC algorithm using the original sample size, global effective sample size, and local effective sample size respectively (the same for Rank PC). We first analyze how the effective sample size helps in causal discovery. Figure 1 shows that there is just a slight decrease in TPR for both CoPC and RPC, when SS is replaced with GESS or LESS. However, a big improvement in FPR appears under both MCAR and MAR, which is especially prominent for RPC. Therefore, w.r.t. the overall metric SHD, the PC algorithms with GESS or LESS perform better than with SS. Also, we notice that LESS can yield more accurate results than GESS: indistinguishable TPR, but better FPR and SHD. We conclude that: 1) compared to the sample size, the usage of an effective sample size (both GESS and LESS) significantly reduces the number of false positives, which thus leads to a better CPDAG; 2) LESS is a better choice in the conditional independence tests. Also, we notice that CoPC outperforms RPC overall for MCAR, which is even more significant for MAR. This is because CoPC can more accurately learn the correlation matrix from incomplete data and is still unbiased under MAR.

In order to evaluate the performance of Copula PC on mixed data, we simulate data as follows: 1) generate Gaussian data and fill in some missing values (as we did before); 2) discretize 25% variables (randomly chosen) into binary; 3) discretize another 25% into ordinal variables with 5 levels. We then replace the rank correlation in RPC with the so-called Hetcor [25] correlation which tests Pearson correlation between continuous variables, polyserial correlation between continuous and ordinal variables, as well as polychoric correlation between ordinal variables. The resulting algorithm is referred to as the 'Hetcor PC' algorithm (HPC).

The performance of HPC and CoPC for $n = 1000$ on mixed data are shown in Figure 2, for the same experimental setting as in the nonparanormal cases. It shows that HPC and CoPC have similar performance w.r.t. TPR regardless of MCAR and MAR, whereas CoPC works much better than HPC w.r.t. FPR. This is because the effective sample size in

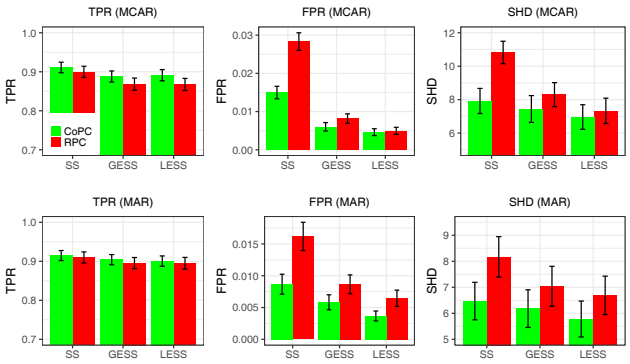

Fig. 1: Performance of Rank PC (RPC) and Copula PC (CoPC) using sample size (SS), global effective sample size (GESS), and local effective sample size (LESS) on nonparanormal data with missing values, showing the mean of TPR, FPR, and SHD over 100 experiments with 95% confidence interval. The two rows represent the results under MCAR and MAR respectively.

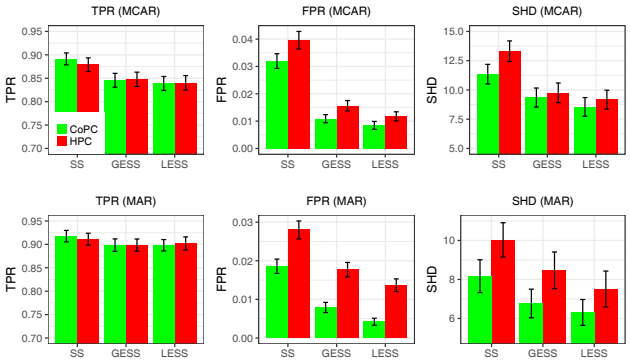

Fig. 2: Performance of Hetcor PC (HPC) and Copula PC (CoPC) on mixed data with missing values, for the same experiments as in Figure 1.

CoPC can account for both mixed data and missing values while HPC corrects only for missing values, which incurs more false positives. As for SHD, CoPC shows an advantage over HPC under MCAR, which becomes much more obvious under MAR because of the biased estimation of the correlations in HPC. Overall, CoPC outperforms HPC on mixed data with missing values, which is more significant when the data are MAR. In addition, we see that the improvement incurred by the usage of an effective sample size still clearly exists for both HPC and CoPC. This is more evident for CoPC in mixed cases than nonparanormal cases, because the effective sample size in CoPC works for both mixed data and missing values.

C. Application to Real-world Data

In this subsection, we compare 'Rank PC' and 'Copula PC' on a dataset of riboflavin production, which is publicly available in the R package **hdi**. It contains 71 continuous-measured observations of 4088 gene expressions. For the ease

TABLE I: Results obtained by various causal discovery algorithms on riboflavinV10 dataset, showing the mean of TPR, FPR, and SHD over 50 repeated experiments with an indication of the number of perfect solutions over these trials.

	TPR	FPR	SHD
MCAR			
RPC + SS	0.63 – 1	0.065 – 2	7.4 – 0
RPC + GESS	0.59 – 0	0.038 – 7	5.9 – 0
RPC + LESS	0.60 – 0	0.038 – 7	5.8 – 0
CoPC + SS	0.81 – 8	0.033 – 15	3.5 – 3
CoPC + GESS	0.76 – 2	0.026 – 18	3.5 – 0
CoPC + LESS	0.71 – 2	0.021 – 22	3.5 – 2
MAR			
RPC + SS	0.83 – 7	0.056 – 4	6.0 – 2
RPC + GESS	0.79 – 5	0.047 – 5	5.5 – 1
RPC + LESS	0.78 – 2	0.036 – 10	5.0 – 2
CoPC + SS	0.94 – 24	0.024 – 19	1.8 – 14
CoPC + GESS	0.92 – 20	0.023 – 19	2.0 – 13
CoPC + LESS	0.91 – 15	0.017 – 24	1.6 – 13

of reproduction, we choose 10 genes with largest empirical variance as our experimental data, denoted by *riboflavinV10* [2]. With all the 71 observations, the 'Rank PC' and 'Copula PC' with significance level 0.05 output the same structure, which is taken as the 'pseudo' ground truth for evaluating resulting graphs of the algorithms on data with missing values.

Then, we fill in missing values to the dataset following the procedure in Section V-A and run our algorithms on the resulting incomplete data. Table I shows the mean of TPR, FPR, and SHD over 50 experiments with an indication of the number of perfect solutions (TPR = 1, FPR = 0, SHD = 0) over these trials. We see that CoPC substantially outperforms RPC under MCAR w.r.t. all the metrics, regardless of SS, GESS, or LESS, which becomes more significant under MAR.

VI. Conclusion

In this paper, we extended the 'Rank PC' algorithm to incomplete data by applying rank correlations to pairwise complete observations and taking the number of pairwise complete observations as an effective sample size. Despite theoretical guarantees, this naive approach only applies to continuous data under MCAR. To make the algorithm more general, we proposed a novel Bayesian approach, in which a Gibbs sampler is designed to draw correlation matrix samples from the posterior distribution given incomplete data. These are then translated into the underlying correlation matrix and the effective sample size for causal discovery. The resulting 'Copula PC' algorithm works for mixed data under MAR, a less restrictive assumption, and even if MAR fails, Bayesian methods like ours can still show strong robustness [12].

For both 'Rank PC' and 'Copula PC', we proposed to replace the sample size with an effective sample size in the tests for conditional independence when that data contains missing values, which significantly improves the performance of the PC algorithm. Especially, we give a local effective sample size to each conditional independence test, since each test only relies on a local structure involving part of the variables (Equation (3)), which makes much sense in particular when some variables contain more missing values than others.

[2]This data and the code is available in https://www.dropbox.com/sh/s8cwbavt9kx14ga/AAB6Y7iNwQjwNinM8dcmox4qa?dl=0

References

[1] J. Pearl, *Causality*. Cambridge university press, 2009.

[2] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[3] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, pp. 294–321, 2012.

[4] N. Harris and M. Drton, "PC algorithm for nonparanormal graphical models," *The Journal of Machine Learning Research*, vol. 14, no. Jan, pp. 3365–3383, 2013.

[5] K. Budhathoki and J. Vreeken, "Causal inference by compression," in *ICDM*. IEEE, 2016, pp. 41–50.

[6] T. Claassen, J. Mooij, and T. Heskes, "Learning sparse causal models is not NP-hard," in *UAI*, 2013, pp. 172–181.

[7] R. Cui, P. Groot, and T. Heskes, "Copula PC algorithm for causal discovery from mixed data," in *ECML PKDD*. Springer, 2016, pp. 377–392.

[8] R. J. Little and D. B. Rubin, "Statistical analysis with missing data," 1987.

[9] K. Mohan and J. Pearl, "Graphical models for recovering probabilistic and causal queries from missing data," in *NIPS*, 2014, pp. 1520–1528.

[10] H. Wang, F. Fazayeli, S. Chatterjee, A. Banerjee, K. Steinhauser, A. Ganguly, K. Bhattacharjee, A. Konar, and A. Nagar, "Gaussian copula precision estimation with missing values." in *AISTATS*, 2014, pp. 978–986.

[11] D. B. Rubin, "Inference and missing data," *Biometrika*, pp. 581–592, 1976.

[12] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art." *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[13] P. D. Hoff, "Extending the rank likelihood for semiparametric copula estimation," *The Annals of Applied Statistics*, pp. 265–283, 2007.

[14] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *Journal of Machine Learning Research*, vol. 2, no. Feb, pp. 445–498, 2002.

[15] T. W. Anderson, *An introduction to multivariate statistical analysis*. John Wiley & Sons, 2003.

[16] M. G. Kendall, "Rank correlation methods." 1948.

[17] W. H. Kruskal, "Ordinal measures of association," *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 814–861, 1958.

[18] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.

[19] J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas, "Bayesian Gaussian copula factor models for mixed data," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 656–665, 2013.

[20] P. D. Hoff, "sbgcop: Semiparametric Bayesian Gaussian copula estimation and imputation," *R package version 0.975*, 2010.

[21] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *The Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 613–636, 2007.

[22] M. Kalisch, M. Mächler, and D. Colombo, "pcalg: Estimation of CPDAG/PAG and causal inference using the IDA algorithm," *URL http://CRAN. R-project. org/package= pcalg.*, 2010.

[23] M. Kolar and E. P. Xing, "Estimating sparse precision matrices from data with missing values," in *ICML*, 2012.

[24] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[25] J. Fox, "Polycor: polychoric and polyserial correlations," *R package version 0.7-5, URL http://CRAN. R-project. org/package= polycor*, 2007.

[26] Q. Cao, Y. Zang, L. Sun, M. Sui, X. Long, Q. Zou, and Y. Wang, "Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study," *Neuroreport*, vol. 17, no. 10, pp. 1033–1036, 2006.