Copula PC Algorithm for Causal Discovery from Mixed Data

Ruifei Cui (✉), Perry Groot, and Tom Heskes

Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands
{R.Cui, Perry.Groot, T.Heskes}@science.ru.nl

Abstract. We propose the 'Copula PC' algorithm for causal discovery from a combination of continuous and discrete data, assumed to be drawn from a Gaussian copula model. It is based on a two-step approach. The first step applies Gibbs sampling on rank-based data to obtain samples of correlation matrices. These are then translated into an average correlation matrix and an effective number of data points, which in the second step are input to the standard PC algorithm for causal discovery. A stable version naturally arises when rerunning the PC algorithm on different Gibbs samples. Our 'Copula PC' algorithm extends the 'Rank PC' algorithm, which has been designed for Gaussian copula models for purely continuous data. In simulations, 'Copula PC' indeed outperforms 'Rank PC' in cases with mixed variables, in particular for larger numbers of data points, at the expense of a slight increase in computation time.

Keywords: causal discovery, Gaussian copula, mixed data, extended rank likelihood

1 Introduction

Causal discovery, or causal structure learning [23], aims to find an underlying directed acyclic graph (DAG), which represents direct causal relations between variables. It is a very popular approach for multivariate data analysis and therefore is widely studied in the past few years, resulting in lots of algorithms. The PC [27, 28] algorithm can be considered the reference causal discovery algorithm. It makes use of conditional independence tests to build the underlying DAG from observations. Starting from a complete undirected graph, the PC algorithm removes edges recursively according to the outcome of the conditional independence tests. This procedure yields an undirected graph, also called the skeleton. After applying various edge orientation rules, it finally gives back a partially directed graph to represent the underlying DAGs.

One advantage of the PC algorithm is that it is computationally feasible for sparse graphs even with thousands of variables. Therefore, it is widely used in high-dimensional settings, generating a variety of applications [20], [29]. Also, open-source software is available like *pcalg* [17] and the Tetrad project [25].

When applied to Gaussian models, the PC algorithm tests conditional independence using partial correlation based on Pearson correlations between variables: when the joint distribution is a multivariate Gaussian, pairwise conditional independence is equivalent to the vanishing of the corresponding partial correlation [18]. Following [12], we will refer to the PC algorithm for Gaussian models as the 'Pearson PC' algorithm. As input it takes the correlation matrix of the observed data and the number of data points. The number of data points is needed for the conditional independence tests: the higher the number of data points, the more reliable the observed correlation matrix as an estimate of the (unknown) true correlation matrix, and the more easily the null hypothesis of conditional independence (given the same value for the partial correlation and the significance level) gets rejected. Under relatively mild assumptions regarding the sparseness of the true underlying DAG, the 'Pearson PC' algorithm shows uniform consistency [16].

Harris and Drton [12] extend the PC algorithm to non-parametric Gaussian (nonparanormal) models, i.e., continuous data assumed to be generated from a Gaussian copula model. They propose to apply the standard PC algorithm, but then replacing the Pearson correlation matrix with rank-based measures of correlation. The so-called 'Rank PC' (RPC) algorithm works as well as the 'Pearson PC' algorithm on normal data and much better on non-normal data, and is shown to be uniformly consistent in high-dimensional settings.

In this paper, we aim to generalize the 'Pearson PC' and 'Rank PC' algorithm to Gaussian copula models that can also handle binary and ordinal variables. The 'Rank PC' algorithm is explicitly limited to the continuous situation, where ties appear with probability zero, making ranks well-defined. In the presence of binary and ordinal variables, ties make the rank correlations between observed variables different from those between the corresponding latent variables in the Gaussian copula setting. Ignorance of this difference typically leads to underestimates of the (absolute) correlations [13].

It is tempting to follow a similar two-step approach as for 'Rank PC': first estimate the correlation matrix in the latent space and then use this as input to the standard PC algorithm. This, however, is not as straightforward as it may seem, for two reasons. First, because of the ties, estimating the correlation matrix of Gaussian copula models for mixed data is considerably more complicated. Second, the ties imply a loss of information, which makes that our estimate of the correlation matrix will tend to be less reliable than in the fully continuous case, which should be accounted for when applying the conditional independence tests in the PC algorithm.

To solve both issues, we propose to make use of a Gibbs sampling procedure, specifically the one derived by Hoff [13] based on the so-called extended rank likelihood. This procedure is relatively straightforward and easy to implement (see the code in the Appendix of [13]). For purely Gaussian data, the correlation matrix samples follow a specific kind of inverse-Wishart distribution [3], which we refer to as the projected inverse-Wishart distribution. Projected inverse-Wishart distributions are characterized by two parameters: the scale matrix and the

degrees of freedom; the former relates to the average correlation matrix and the latter to the number of data points. As we will show, under the projected inverse-Wishart, the variance of each off-diagonal element of the correlation matrix is an approximate function of its expectation and the degrees of freedom: the more degrees of freedom, the smaller the variance. The idea is now to estimate the scale matrix and degrees of freedom from the Gibbs samples of more general Gaussian copula models on mixed data, as if they were also drawn from a projected inverse-Wishart distribution. The scale matrix is translated into a correlation matrix and the degrees of freedom into a so-called 'effective number of data points', to take into account the reliability of our estimate of the correlation matrix. These are then input to the standard PC algorithm for causal discovery.

We refer to our two-step procedure as the 'Copula PC' (CoPC) algorithm. We also derive a stable version, referred to as 'Stable Copula PC' (SCPC), which runs PC repeatedly on a number of Gibbs samples. Experimental results show that both CoPC and SCPC outperform the current 'Rank PC' algorithm in mixed databases with discrete and continuous variables.

The rest of this paper is organized as follows. Section 2 reviews some relevant background information and analyzes issues of existing algorithms in more detail. Section 3 proposes an approximate inference method for the correlation matrix and the effective number of data points based on the projected inverse-Wishart distribution, and then derives the resulting algorithms CoPC and SCPC. Section 4 compares CoPC and SCPC with the 'Rank PC' algorithm on simulated data and provides an illustration on real-world data of ADHD patients. Section 5 gives conclusions and future work.

2 Preliminaries and Problem Analysis

In this section, we first review some necessary background information on causal discovery, then briefly introduce the PC and 'Rank PC' algorithm, and finally analyze the challenges current PC algorithms face for mixed data.

2.1 Causal Structure Learning

A graph $G = (\boldsymbol{V}, \boldsymbol{E})$ consists of a set of vertices $\boldsymbol{V} = \{X_1, \ldots, X_p\}$, representing random variables, and a set of edges $\boldsymbol{E}$, representing relations between pairs of variables. A graph is *directed* if it only contains directed edges while it is *undirected* if it only contains undirected edges. Graphs containing both directed and undirected edges are called *partially directed* graphs. A graph with no directed cycles, e.g., $X_i \to X_j \to X_i$ is *acyclic*. A graph which is both directed and acyclic is a *Directed Acyclic Graph* (DAG). If there is an edge from X_i to X_j, X_i is the parent of X_j. The set of parents of X_j in graph G is denoted by $pa(G, X_j)$.

A multivariate probability distribution P over variables $\boldsymbol{V} = \{X_1, \ldots, X_p\}$ is said to *factorize* according to a DAG $G = (\boldsymbol{V}, \boldsymbol{E})$, if the joint probability density function of P can be written as the product of the conditional densities of each variable given its parents in G, i.e., $f(X_1, \ldots, X_p) = \prod_{i=1}^{p} f(X_i | pa(G, X_i))$. If

this condition holds, we can read off conditional independence relationships in distribution P from the DAG via a graphical criterion called *d-separation* [24]. *D-separation* implies that each variable is independent of its non-descendants given its parents.

Several DAGs may, via *d-separation*, correspond to the same set of conditional independencies. Such DAGs form a Markov equivalence class, which can be uniquely represented by a completed partially directed acyclic graph (CPDAG) [6]. Arcs in a CPDAG indicate a cause-effect relation between variables since the same arc occurs in all members of the CPDAG. Undirected edges $X_i - X_j$ in a CPDAG indicate that some of its members contain an arc $X_i \to X_j$ whereas other members contain an arc $X_j \to X_i$. The aim of causal discovery is to learn the Markov equivalence class of a DAG $G = (\boldsymbol{V}, \boldsymbol{E})$ from n i.i.d. observations of $\boldsymbol{V}$.

2.2 PC Algorithm and Rank PC Algorithm

The PC algorithm starts from a complete undirected graph, and then removes edges recursively according to conditional independencies yielding a partially connected undirected graph called the skeleton, after which some orientation rules are applied to direct as many edeges as possible, resulting in a completed partially directed acyclic graph, i.e. the underlying CPDAG.

During the process, testing conditional independence plays the most important role. The PC algorithm uses partial correlation, denoted by $\rho_{uv|S}$, to do it. The correlation matrix from independent observations of a random vector $\boldsymbol{Z}$ can be used to estimate $\rho_{uv|S}$ [2]. Then, classical decision theory is applied to judge conditional independencies using significance level α,

$$Z_u \perp\!\!\!\perp Z_v | Z_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{uv|S}}{1 - \hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2), \qquad (1)$$

where $u \neq v$ and $S \subseteq \{1, \ldots, p\} \backslash \{u, v\}$. Thus in order to run the PC algorithm we need the correlation matrix corresponding to the data to estimate $\hat{\rho}_{uv|S}$ and the number of observations n.

The PC algorithm has been extended to a broader class of Gaussian copula by using rank correlations to replace Pearson correlations, resulting in the 'Rank PC' algorithm [12]. Rank correlations, typically Spearman's ρ and kendall's τ, only consider the ranks among observations, ignoring the actual variables.

Definition 1 (Gaussian Copula Model). *Consider two random vectors* $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ *and* $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$*, satisfying the conditions* $\boldsymbol{Z} \sim \mathcal{N}(0, C)$ *and* $Y_i = F_i^{-1}(\Phi(Z_i))$ *for* $i = 1, \ldots, p$ *where* C *denotes the correlation matrix of* $\boldsymbol{Z}$ *and* ${F_i}^{-1}(t) = \inf\{y : F_i(y) \geq t\}$ *is the pseudo-inverse of a cumulative distribution function* F_i*. Then this model is called* Gaussian copula model *with correlation matrix* C *and univariate margins* F_i*.*

In the Gaussian copula model, when all margins are continuous, ties occur with zero probability making ranks well-defined. For such so-called nonparanormal models, the sample correlations among ranks can naturally be used as an

estimator for the Pearson correlation in the latent space. In this nonparanormal setting, RPC has been shown to perform well [12], [19].

2.3 Challenges for Mixed Data

RPC works well on continuous data, because tied observations occur with probability zero. In the presence of discrete margins, however, the estimator used in RPC is no longer consistent because of the tied observations. In this case, standard rank-based correlation will be different from the true correlation in latent space [13], typically underestimating it. Hence, our first challenge is to estimate the underlying C efficiently and consistently from mixed data.

A second challenge concerns the information loss incurred by discrete variables. Specifically, simply setting n in Equation (1) to the number of data points can lead to an underestimate of the p-values provided by the conditional independence tests. To solve this problem, we introduce the notion of an effective number of data points.

3 Approximate Inference and Copula PC Algorithm

In this section, we introduce an approximate inference approach for the underlying correlation matrix and the effective number of data points from mixed data. Subsection 3.1 introduces the projected inverse-Wishart distribution and its application to Gaussian models. Subsection 3.2 discusses how to obtain correlation matrix samples from mixed data using a Gibbs sampling procedure. Subsection 3.3 shows how to use these samples to estimate the two parameters of the projected inverse-Wishart distribution: the scale matrix (as the underlying correlation matrix) and the degrees of freedom (as the effective number of data points). Subsection 3.4 gives the resulting Copula PC algorithm and the Stable Copula PC algorithm.

3.1 Projected Inverse-Wishart Distribution

Priors on correlation matrices are typically derived by choosing the inverse-Wishart distribution, denoted by $\mathcal{W}^{-1}(\Sigma; \Psi_0, \nu)$, as a prior on covariance matrices and then turning the covariance matrices into a correlation matrix to end up with an implied distribution on the correlation matrix. We choose Σ from $\mathcal{W}^{-1}(\Sigma; \Psi_0, \nu)$ and write

$$P(C) = \mathcal{PW}^{-1}(C; \Psi_0, \nu) \tag{2}$$

where $C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$ for $\forall\ i, j$. Since many covariance matrices possibly correspond to the same correlation matrix, the above process can be considered as a projection from covariance matrices to a correlation matrix. Therefore, we refer to this distribution on correlation matrix C as a projected inverse-Wishart distribution.

For Gaussian models, the projected inverse-Wishart distribution gives exact inference [21]. Specifically, given data $\boldsymbol{Z} = (\boldsymbol{z_1}, \ldots, \boldsymbol{z_n})$, the posterior reads

$$P(\Sigma|\boldsymbol{Z}) = \mathcal{W}^{-1}(\Sigma; \Psi_0 + \Psi, \nu + n) \text{ and } P(C|\boldsymbol{Z}) = \mathcal{PW}^{-1}(C; \Psi_0 + \Psi, \nu + n)\,,$$

with $\Psi = \boldsymbol{Z}^T\boldsymbol{Z}$. Also, the projected inverse-Wishart is scale invariant [3], [14], in the sense that we can make the posterior distribution on correlation matrices independent of the scale of the data by choosing $\Psi_0 = \mathbb{0}$, or perhaps better, $\Psi_0 = \epsilon\mathbb{1}$ in the limit $\epsilon \downarrow 0$.

Summarizing, we consider the prior distribution

$$P(\Sigma) = \mathcal{W}^{-1}(\Sigma; \epsilon\mathbb{1}, p+1) \text{ in the limit } \epsilon \downarrow 0,$$

which in fact boils down to the well-known improper Jeffreys prior [32]:

$$P(\Sigma) \propto \|\Sigma\|^{-(p+1)}\,.$$

For Gaussian copula models, although there is no analytical expression, we still expect that the posterior $P(C|\boldsymbol{Y})$ can be approximated through a projected inverse-Wishart distribution, i.e., $P(C|\boldsymbol{Y}) \approx \mathcal{PW}^{-1}(C; \Psi, \nu)$ for some Ψ and ν.

3.2 Gibbs Sampler Based on Extended Rank Likelihood

Hoff [13] describes an elegant procedure to obtain samples from $P(C|\boldsymbol{Y})$ for a Gaussian copula model. The essence is that we only consider the ranks among observations, hence the name extended rank likelihood, ignoring the actual variables. Since the cumulative distribution functions $F_i(Y_i)$ are non-decreasing, observing $y_{i_1,j} < y_{i_2,j}$ implies that $z_{i_1,j} < z_{i_2,j}$, where $y_{i_1,j}$ denotes the i_1^{th} observation of the j^{th} component of random vector $\boldsymbol{Y}$, To be precise, observing $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ tells us that $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ must lie in the set

$$\left\{\boldsymbol{Z} \in \mathbb{R}^{n\times p} : \max\left\{z_{k,j} : y_{k,j} < y_{i,j}\right\} < z_{i,j} < \min\left\{z_{k,j} : y_{i,j} < y_{k,j}\right\}\right\}.$$

Strong posterior consistency for C under the extended rank likelihood has been proved in the situation with both discrete and continuous marginal distribution functions [22].

An off-the-shelf sampling algorithm based on the extended rank likelihood is full Gibbs sampling [13]. The code of this sampling algorithm is provided in the Appendix of [13]. In this algorithm, each component of $\boldsymbol{Z}$ is initialized according to the rank information of the corresponding component of $\boldsymbol{Y}$, after which each component is resampled alternatively. Here we propose a slight modification by just resampling the discrete components instead of all of them. Experimental tests reveal that the results of this faster sampling approach are indistinguishable from Hoff's original Gibbs sampler. Although this modification is quite straightforward, it significantly reduces computation time because sampling continuous variables is far more time-consuming than sampling discrete ones in Hoff's Gibbs sampler. We will refer to this modified sampling algorithm as *SamplingAlgo*.

So, given the observed data $\boldsymbol{Y}$, samples on the underlying correlation matrix, denoted by $\{C^{(1)}, \ldots, C^{(m)}\}$, can be obtained using *SamplingAlgo*.

3.3 Estimation of the Correlation Matrix and the Effective Number of Data Points

This subsection aims to estimate the underlying correlation matrix and the effective number of data points from the obtained samples.

Theorem 1 suggests a procedure to estimate the parameters Ψ and ν from samples of a projected inverse-Wishart distribution $\mathcal{PW}^{-1}(C; \Psi, \nu)$.

Theorem 1. *If the correlation matrix C follows a projected inverse-Wishart distribution with parameters Ψ ($\Psi_{ii} = 1$) and ν, i.e.,*

$$P(C) = \mathcal{PW}^{-1}(C; \Psi, \nu),$$

then for each off-diagonal element $C_{ij} (i \neq j)$ and large ν, we have

$$\mathrm{E}\left[C_{ij}\right] \approx \Psi_{ij} \quad \textit{and} \quad \mathrm{Var}\left[C_{ij}\right] \approx \frac{(1 - (\Psi_{ij})^2)^2}{\nu} \,.$$

The proof is given in the Appendix.

According to Theorem 1, the mean over samples of C is an excellent approximation of Ψ. As for ν, we have,

$$\nu \approx \frac{(1 - (\mathrm{E}\left[C_{ij}\right])^2)^2}{\mathrm{Var}\left[C_{ij}\right]} \,. \tag{3}$$

The idea now is to apply the same estimates, as if the samples obtained by Gibbs sampling the Gaussian copula model on mixed data also (approximately) follow a projected inverse-Wishart distribution. Specifically, for the effective number of data points $\hat{n}$, we propose to take the average over all $p(p-1)/2$ estimates on ν that can be computed by applying (3) to each upper triangular element of a p-dimensional correlation matrix C.

3.4 Copula PC Algorithm and Stable Copula PC Algorithm

Now, we turn the previous results into a working algorithm. The two key input arguments of the 'Pearson PC' algorithm are the correlation matrix and the number of data points. In the general Gaussian copula model, we take the mean over $\{C^{(1)}, \ldots, C^{(m)}\}$ and the mean over $p(p-1)/2$ estimates on ν as the two arguments respectively, resulting in the Copula PC algorithm.

Next, we introduce a stable version of the Copula PC algorithm. We take l instances from all the m samples. For each instance, a corresponding graph can be obtained via the 'Pearson PC' algorithm using the earlier estimated effective number of data points, by which a collection of l graphs can be generated, denoted by $\{\widetilde{G}_1, \ldots, \widetilde{G}_l\}$. We keep those edge marks that emerge with a probability higher than a pre-defined threshold β and remove the others, leading to a resulting graph. Since this resulting graph seemingly contains only 'stable' edge

Algorithm 1 Copula PC algorithm and its stable version

Input: Observations $\boldsymbol{Y}$, Initialized parameters m, l, β
Output: Causal graph G_c by CoPC, G_s by SCPC
$C^{(1)}, \ldots, C^{(m)} = SamplingAlgo(\boldsymbol{Y})$
for all C_{ij} with $i < j$ (upper triangular elements) **do**
    Compute and store $\nu_k = \frac{(1-(\mathrm{E}[C_{ij}])^2)^2}{\mathrm{Var}[C_{ij}]}$ ▷ Equation (3)
end for
$\hat{n}$ = the average over $\{\nu_1, \ldots, \nu_{p(p-1)/2}\}$
if CoPC **then** ▷ procedures for CoPC
    $\hat{C} = \frac{1}{m}\sum_{j=1}^{m} C^{(j)}$
    $G_c = pc(\hat{C}, \hat{n})$ ▷ the 'Pearson PC' algorithm
else ▷ procedures for SCPC
    Choose l ($l < m$) instances from $C^{(1)}, \ldots, C^{(m)}$
    for $i = 1 : l$ **do**
        Compute and store $\widetilde{G}_i = pc(C^{(l_i)}, \hat{n})$
    end for
    for all edge marks **do**
        e = the number of graphs containing the current edge mark
        if $e/l > \beta$ **then**
            keep the edge mark
        end if
        G_s = all kept edge marks among $\{\widetilde{G}_1, \ldots, \widetilde{G}_l\}$.
    end for
end if

marks, we call this method stable Copula PC algorithm (SCPC). The size of l has a linear influence on running time because choosing l means the 'Pearson PC' algorithm would run l times. As for β, a small value means keeping more edge marks and vice versa. The Copula PC algorithm and its stable version are summarized in Algorithm 1.

4 Experiments

In this section, we first verify the property of the projected inverse-Wishart distribution described by Equation (3) and check whether it still holds in the presence of discrete variables. Then, we compare the proposed CoPC and SCPC with the 'Rank PC' algorithm on simulated data and give an illustration on real-world data of ADHD patients.

Following Kalisch and Bühlmann [16], we simulate random DAGs and draw samples from the distributions faithful to them. Firstly, we generate an adjacency matrix A, whose entries are zero or in the interval $[0.1, 1]$. There exists a directed edge from i to j in the corresponding DAG, if $i < j$ and $A_{ji} \neq 0$. The DAGs generated in this way have the property $\mathrm{E}(N_i) = s(p-1)$, where N_i is the number of neighbors of node i, and s is the probability that there is an edge between any two nodes, called the sparseness parameter. Then, the samples of

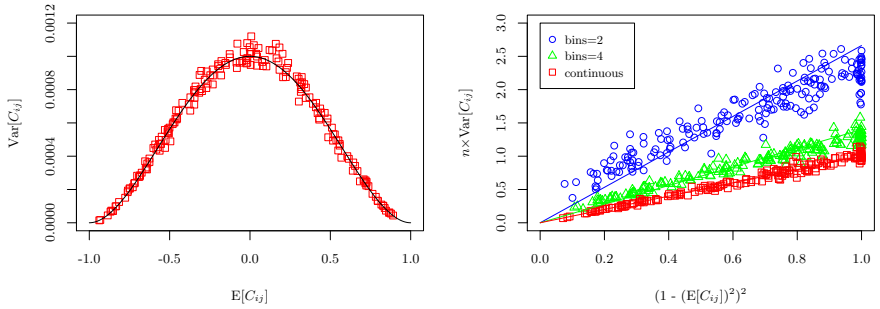

Fig. 1: The relationship between the expectation and the variance of the elements of sampled correlation matrices. Left panel: The samples are drawn from a given projected inverse-Wishart distribution. Right panel: The samples are drawn via *SamplingAlgo*, with circles for binary cases, triangles for ordinal cases with 4 levels, and squares for continuous cases.

a random vector $\boldsymbol{Z}$ are drawn through

$$\boldsymbol{Z} = A\boldsymbol{Z} + \epsilon\,, \tag{4}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ is a vector of independent standard normal random variables. The data generated in this way follow a multivariate Gaussian distribution.

4.1 Estimation for the Effective Number of Data Points

As argued in Subsection 3.3, the expectation and variance of the elements of correlation matrices drawn from a projected inverse-Wishart distribution are strongly related. To check this relationship, we proceed as follows: 1) we generate a random p-dimensional correlation matrix Ψ; 2) we draw 500 samples from a projected inverse-Wishart distribution with parameters Ψ and ν; 3) for each upper triangular element, we plot its variance against its expectation.

The left panel in Figure 1 shows a typical result for $p = 20$ and $\nu = 1000$. We see that almost all pairs are distributed around the theoretical curve (solid line) especially when the expectation is far from zero, which indicates that it is indeed possible to infer ν of a projected inverse-Wishart distribution via the expectation and variance of off-diagonal elements.

Next, we study how our inference method works for estimating $\hat{n}$ in different cases. We first generate n samples of $\boldsymbol{Z}$ using Equation (4) and discretize some of the variables to obtain the simulated samples of the observed random vector $\boldsymbol{Y}$. Then, we run *SamplingAlgo* to get samples of the underlying C. The results for $p = 20$ and $n = 1000$ for different cases are shown in Figure 1 (right panel), where 'bins=2' means that all variables are binary, 'bins=4' means that all variables are ordinal with 4 levels and 'continuous' means that all variables are kept continuous. We take $(1 - (\mathrm{E}\,[C_{ij}])^2)^2$ for the x-axis and $n \times \mathrm{Var}\,[C_{ij}]$ for the y-axis, so that all data points are expected to be distributed around a

straight line with slope $n/\hat{n}$. For purely continuous variables, a straight line with slope 1 gives an almost perfect fit, as expected. For ordinal and binary variables, we still find a clear trend, but mild deviations from a perfect straight line, indicating that the projected inverse-Wishart distribution is a fine, but not perfect approximation of the exact posterior. The stronger the discretization, the larger the slope $n/\hat{n}$ and thus the lower our estimated effective number of data points.

More extensive experiments (not shown) done with different numbers of variables, data points, Gibbs samples and sparseness parameters, reveal that these hardly influence the general picture, as long as the number of data points and the number of Gibbs samples are both at least 100.

4.2 Causal Discovery on Simulations

In this subsection, we compare CoPC and SCPC with the 'Rank PC' [12] algorithm. All computations are implemented in the R-package *pcalg*.

We first generate multivariate normal data (p variables) via Equation (4). After that, 25% of all p variables are discretized into binary variables, and another 25% of them are discretized into ordinal variables with 5 levels. In this way, we simulate the observations of $\boldsymbol{Y}$ which are generated from a Gaussian copula model with both discrete and continuous margins.

Three measures are used to test the performance: 1) percentage of correct edges in the resulting skeleton, usually called true positive rate (TPR); 2) percentage of spurious edges, usually called false positive rate (FPR); 3) Structural Hamming Distance (SHD), counting the number of edge insertions, deletions, and flips in order to transfer the estimated CPDAG into the correct CPDAG [30]. The first two measures are for the skeleton while SHD is for the CPDAG. A smaller SHD indicates better performance.

Next, we compare the performance of three versions of the PC algorithm, RPC, CoPC, and SCPC in terms of TPR, FPR, and SHD. We restrict the significance level to $\alpha = 0.01$, which has been shown to yield the best overall SHD [16]. For CoPC, we drop the first 20 Gibbs samples and save the next 100 samples ($m = 100$). For SCPC, we take $l = 20$ equidistant samples, so $\{C^{(1)}, C^{(6)}, \ldots, C^{(96)}\}$, and choose β such that the TPR for SCPC is more or less equal to that of RPC, which amounts to $\beta = 0.4$ for sparse graphs with 10 nodes, $\beta = 0.45$ for sparse graphs with 50 nodes, and $\beta = 0.3$ for dense graphs. The remaining parameters are set as follows: $p \in \{10, 50\}$, $n \in \{500, 1000, 2000, 5000\}$, and $\mathrm{E}[N] \in \{2\ (Sparse), 5\ (Dense)\}$.

The comparative results in Figure 2 (10 nodes) and Figure 3 (50 nodes) provide the mean over 100 repeated experiments and errorbars representing 95% confidence intervals. First, for sparse graphs (both small and large graphs), the three algorithms get nearly the same results w.r.t. TPR, but CoPC and SCPC show a large advantage over RPC w.r.t. FPR and SHD except SCPC with large graphs, which becomes more prominent with increasing sample size. Second, for dense graphs, the advantage of CoPC and SCPC over RPC still exists w.r.t. FPR, although seemingly CoPC performs a little worse than SCPC and RPC w.r.t. TPR. Third, we note that the performance of RPC deteriorates seriously w.r.t.

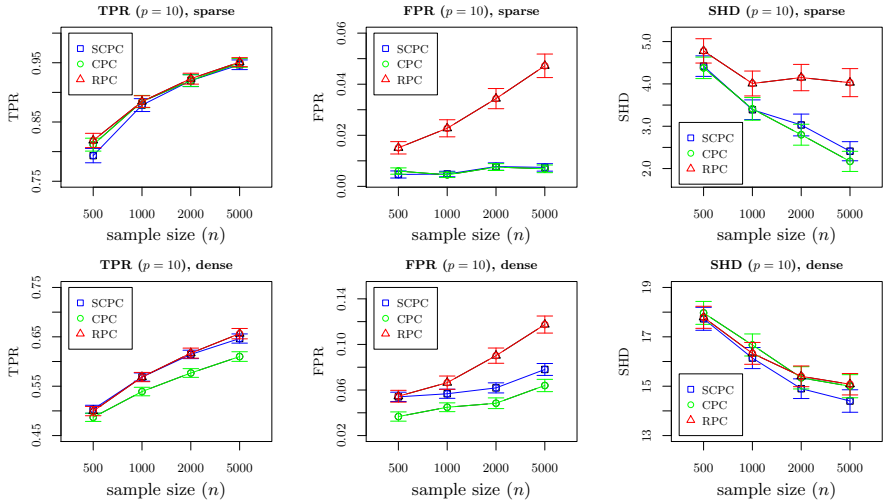

Fig. 2: Performance of Rank PC, Copula PC, and Stable Copula PC for 10 nodes, showing the mean of TPR, FPR, and SHD over 100 experiments together with 95% confidence intervals. The first row represents the results with sparse graphs ($\mathrm{E}[N] = 2$) while the second row represents those with dense graphs ($\mathrm{E}[N] = 5$).

FPR with the increase in sample size, while CoPC and SCPC are very stable. Apparently, using sample size as the effective number of data points, RPC incurs more false positives especially for larger sample sizes. Overall, CoPC and SCPC clearly outperform RPC, especially in the sparse cases with larger sample sizes.

4.3 Application to Real-World Data

In this subsection, we give an illustration on a real-world dataset on phenotypic information about children with Attention Deficit Hyperactivity Disorder (ADHD) [5]. It contains 23 variables for 245 subjects. We focus on nine variables as in [26], but keep all subjects with missing values since these are easily handled by the Gibbs sampler. The nine variables considered are: gender (G), attention deficit level (AD), hyperactivity/impulsivity level (HI), verbal IQ (VIQ), performance IQ (PIQ), full IQ (FIQ), aggressive behavior (Agg), medication status (Med), handedness (HN), where four of them (G, Agg, Med, HN) are binary.

We run CoPC and SCPC ($l = 30$, $\beta = 0.4$) on the dataset and consider prior knowledge that no variable can cause gender. The resulting graphs are shown in Figure 4. The graphs suggest that gender has an effect on attention deficit level, which then causes hyperactivity/impulsivity level. This point has been confirmed by many studies [4], [31]. It is common that AD and Agg cause patients to take medicine. Also, VIQ, PIQ, and FIQ are connected to each other by bi-directed edges. This indicates that the causal sufficiency assumption is violated, i.e., that there should be a latent common cause related to IQ, as also suggested in [26].

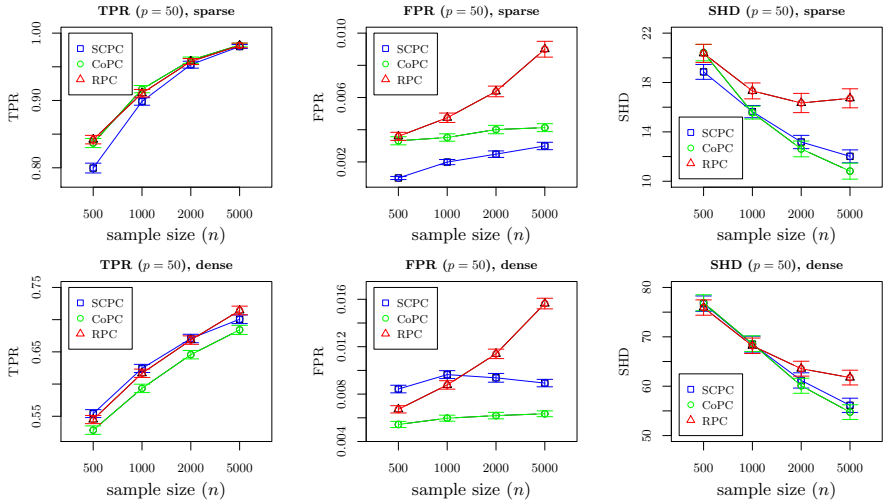

Fig. 3: Performance of Rank PC, Copula PC, and Stable Copula PC for 50 nodes, showing the mean of TPR, FPR, and SHD over 100 experiments together with 95% confidence intervals.

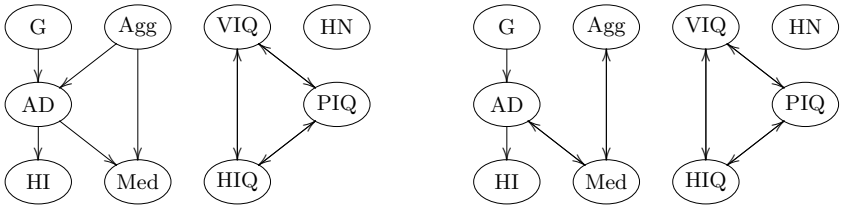

Fig. 4: The resulting graphs by CoPC (left panel) and SCPC (right panel) on ADHD dataset.

5 Conclusions and Future Work

In this paper we introduced a novel two-step approach for estimating the causal structure underlying a Gaussian copula model on mixed data. The essence is to estimate the correlation matrix in the latent space, which can then be given to any causal discovery algorithm to search for its underlying structure. Ties between the discretized observations incur information loss, making the estimate of correlation matrix less reliable than in fully continuous cases. For this, we introduced the notion of 'effective number of data points' that can be estimated from the expectation and variance of the correlation matrix elements. Our approach, based on ranks and correlation matrices, is fully scale invariant and has a natural uninformative setting when choosing a uniform distribution over pairwise correlations, which can be adjusted to account for different assumptions.

We like to think of our two-step approach as a general principle, where for each of the two steps one could plug in one's favorite choice: e.g., a different MCMC method [15] or a MAP approach along the lines of [1] for estimating the correlation matrix and its reliability, and another method, like FCI [28] or BCCD [7], for causal structure learning. Having generated samples, running the PC algorithm several times to gain an insight into the reliability of structure estimates is an obvious thing to do. Similar procedures have been proposed, e.g., by bootstrapping the original dataset [8], [10]. In our simulations, the Gibbs sampler appears to converge quite fast, which makes Gibbs sampling cheap compared to running the PC algorithm, in particular for models with many variables. Our choice to only resample the discrete random variables and not the continuous ones, here also helps. Being fully Bayesian about structure learning as well may be very nice in theory [11], but is computationally infeasible in practice for any reasonable number of variables. Altogether, our Bayesian approach to sample correlation matrices in combination with a more frequentist approach towards structure learning attempts to combine the best of both worlds.

Our methods require the setting of just a few parameters: the significance level α to be used in the PC algorithm (typically 0.01 or 0.05), the number of Gibbs samples and burn-in (the more, the better), and for SCPC, the number of instances l in the ensemble (the more, the better), and the threshold β (the higher, the more conservative).

Our estimate of the 'effective number of data points' appears to work nicely in practice, but can and perhaps should be further improved. Instead of considering the variance of the elements of the correlation matrix, one may come up with another, more direct estimate, for example the entropy of the distribution and translate that into an effect number of data points. Preliminary attempts in that direction failed by being typically much less robust than the one described in this paper. Our current estimate gives a single, global value for the effective number of data points. Future work may consider estimating a different value for each conditional independence test, since each test only relies on a local structure, involving only part of the variables. Such estimates then can be integrated into the causal discovery algorithm itself. Another line of future research concerns the theoretical analysis of CoPC and SCPC, where it can be studied to what extent and under which conditions consistency can be proven. Our conjecture here is that consistency of our two-step procedure follows from the proven consistency of the two separate steps: Gibbs sampling to estimate the correct correlation matrix C [22] and the PC algorithm to arrive at the correct causal structure [16].

Appendix: Proof of Theorem 1

Consider partitioning the matrix Σ and Ψ as

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \quad \text{and} \quad \Psi = \begin{bmatrix} \Psi_{aa} & \Psi_{ab} \\ \Psi_{ba} & \Psi_{bb} \end{bmatrix} .$$

Then, if $P(\Sigma) = \mathcal{W}^{-1}(\Sigma; \Psi, \nu)$, we have

$$\begin{aligned} P(\Sigma_{aa}) &= \mathcal{W}^{-1}(\Sigma_{aa}; \Psi_{aa}, \nu - \dim(b)) \,, \\ P(\Sigma_{bb|a}) &= \mathcal{W}^{-1}(\Sigma_{bb|a}; \Psi_{bb|a}, \nu) \,, \\ P(\Sigma_{aa}^{-1}\Sigma_{ab}|\Sigma_{bb|a}) &= \mathcal{N}(\Sigma_{aa}^{-1}\Sigma_{ab}; \Psi_{aa}^{-1}\Psi_{ab}, \Sigma_{bb|a} \otimes \Psi_{aa}^{-1}) \,, \end{aligned} \tag{5}$$

where $\dim(b)$ is the dimension of Σ_{bb} and $\Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$ [9].

Without loss of generality, we restrict our analysis to a two-dimensional system and suppose that we draw

$$\Sigma \sim \mathcal{W}^{-1}\left(\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \nu\right) \,.$$

Then, according to (5), we have

$$\Sigma_{11} \sim \mathcal{W}^{-1}(1, \nu - 1) \,, \quad \Sigma_{22|1} \sim \mathcal{W}^{-1}(1 - \rho^2, \nu) \,, \quad \Sigma_{11}^{-1}\Sigma_{12}|\Sigma_{22|1} \sim \mathcal{N}(\rho, \Sigma_{22|1}) \,.$$

Rewriting the resulting $\hat{\rho}$ in terms of these variables, we obtain

$$\hat{\rho} = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}\sqrt{\Sigma_{22}}} = \frac{(\Sigma_{11}^{-1}\Sigma_{12})\sqrt{\Sigma_{11}}}{\sqrt{\Sigma_{22|1} + \Sigma_{11}(\Sigma_{11}^{-1}\Sigma_{12})^2}} \,. \tag{6}$$

Since for large ν,

$$\mathrm{E}\left[\Sigma_{11}\right] = \frac{1}{\nu - 3} \approx \frac{1}{\nu} \,, \quad \mathrm{E}\left[\Sigma_{22|1}\right] = \frac{1 - \rho^2}{(\nu - 2)} \approx \frac{1 - \rho^2}{\nu} \,,$$

$$\mathrm{Var}\left[\Sigma_{11}\right] = \frac{2}{(\nu - 3)^2(\nu - 5)} \approx \frac{2}{\nu^3} \,, \quad \mathrm{Var}\left[\Sigma_{22|1}\right] = \frac{2(1 - \rho^2)^2}{(\nu - 2)^2(\nu - 4)} \approx \frac{2(1 - \rho^2)^2}{\nu^3}.$$

we can approximate,

$$\begin{aligned} \Sigma_{11} \approx \frac{1}{\nu}\left(1 + \sqrt{\frac{2}{\nu}}x\right), \Sigma_{22|1} \approx \frac{1 - \rho^2}{\nu}\left(1 + \sqrt{\frac{2}{\nu}}y\right), \\ \Sigma_{11}^{-1}\Sigma_{12} \approx \rho + \sqrt{\frac{1 - \rho^2}{\nu}}z \,, \end{aligned} \tag{7}$$

where x, y, and z are independent random variables, all with mean zero and unit variance. Indeed, for large ν, all noise terms scale with $\sqrt{1/\nu}$ relative to the mean, and can hence be ignored when computing the expectation, to yield, as expected,

$$\mathrm{E}\left[\hat{\rho}\right] \approx \rho \,. \tag{8}$$

To estimate the variance, we substitute (7) into (6), and compute (in leading order, and evaluated for $x = y = z = 0$),

$$\frac{\partial\hat{\rho}}{\partial x} \approx \rho(1 - \rho^2)\sqrt{\frac{1}{2\nu}} \,, \quad \frac{\partial\hat{\rho}}{\partial y} \approx \rho(1 - \rho^2)\sqrt{\frac{1}{2\nu}} \,, \quad \frac{\partial\hat{\rho}}{\partial z} \approx (1 - \rho^2)^{3/2}\sqrt{\frac{1}{\nu}} \,,$$

yielding the variance

$$\mathrm{Var}\left[\hat{\rho}\right] = \left(\frac{\partial\hat{\rho}}{\partial x}\right)^2 + \left(\frac{\partial\hat{\rho}}{\partial y}\right)^2 + \left(\frac{\partial\hat{\rho}}{\partial z}\right)^2 \approx \frac{(1 - \rho^2)^2}{\nu} \,. \tag{9}$$

References

1. Abegaz, F., Wit, E.: Penalized EM algorithm and copula skeptic graphical models for inferring networks for mixed variables. arXiv preprint arXiv:1401.5264 (2014)
2. Anderson, T.W.: An introduction to multivariate statistical analysis. John Wiley & Sons (2003)
3. Barnard, J., McCulloch, R., Meng, X.L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica pp. 1281–1311 (2000)
4. Bauermeister, J.J., Shrout, P.E., Chávez, L., Rubio-Stipec, M., Ramírez, R., Padilla, L., Anderson, A., García, P., Canino, G.: ADHD and gender: are risks and sequela of ADHD the same for boys and girls? Journal of Child Psychology and Psychiatry 48(8), 831–839 (2007)
5. Cao, Q., Zang, Y., Sun, L., Sui, M., Long, X., Zou, Q., Wang, Y.: Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study. Neuroreport 17(10), 1033–1036 (2006)
6. Chickering, D.M.: Learning equivalence classes of Bayesian-network structures. The Journal of Machine Learning Research 2, 445–498 (2002)
7. Claassen, T., Heskes, T.: A Bayesian approach to constraint based causal inference. arXiv preprint arXiv:1210.4866 (2012)
8. Dai, H., Li, G., Zhou, Z.: Ensembling mml causal discovery. In: Advances in Knowledge Discovery and Data Mining, pp. 260–271. Springer (2004)
9. Eaton, M.L.: Chapter 8: The wishart distribution. In: Multivariate Statistics, Lecture Notes–Monograph Series, vol. 53, pp. 302–333. Institute of Mathematical Statistics, Beachwood, Ohio, USA (2007)
10. Entner, D., Hoyer, P.O.: On causal discovery from time series data using FCI. Probabilistic graphical models pp. 121–128 (2010)
11. Friedman, N., Koller, D.: Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine learning 50(1-2), 95–125 (2003)
12. Harris, N., Drton, M.: PC algorithm for nonparanormal graphical models. The Journal of Machine Learning Research 14(1), 3365–3383 (2013)
13. Hoff, P.D.: Extending the rank likelihood for semiparametric copula estimation. The Annals of Applied Statistics pp. 265–283 (2007)
14. Huang, A., Wand, M.P., et al.: Simple marginally noninformative prior distributions for covariance matrices. Bayesian Analysis 8(2), 439–452 (2013)
15. Kalaitzis, A., Silva, R.: Flexible sampling of discrete data correlations without the marginal distributions. In: Advances in Neural Information Processing Systems. pp. 2517–2525 (2013)
16. Kalisch, M., Bühlmann, P.: Estimating high-dimensional directed acyclic graphs with the PC-algorithm. The Journal of Machine Learning Research 8, 613–636 (2007)
17. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., Bühlmann, P.: Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47(11), 1–26 (2012)
18. Lauritzen, S.L.: Graphical models. Clarendon Press (1996)
19. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics pp. 2293–2326 (2012)

20. Maathuis, M.H., Colombo, D., Kalisch, M., Bühlmann, P.: Predicting causal effects in large-scale systems from observational data. Nature Methods 7(4), 247–248 (2010)
21. Murphy, K.P.: Conjugate Bayesian analysis of the Gaussian distribution. def $1(2\sigma 2)$, 16 (2007)
22. Murray, J.S., Dunson, D.B., Carin, L., Lucas, J.E.: Bayesian Gaussian copula factor models for mixed data. Journal of the American Statistical Association 108(502), 656–665 (2013)
23. Pearl, J.: Causality. Cambridge university press (2009)
24. Pearl, J., et al.: Causal inference in statistics: An overview. Statistics Surveys 3, 96–146 (2009)
25. Scheines, R., Spirtes, P., Glymour, C., Meek, C., Richardson, T.: The TETRAD project: Constraint based aids to causal model specification. Multivariate Behavioral Research 33(1), 65–117 (1998)
26. Sokolova, E., Groot, P., Claassen, T., Heskes, T.: Causal discovery from databases with discrete and continuous variables. In: Probabilistic Graphical Models, pp. 442–457. Springer (2014)
27. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction and search. Lecture Notes in Statist 81 (1993)
28. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction, and search. MIT press (2000)
29. Stekhoven, D.J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M.H., Bühlmann, P.: Causal stability ranking. Bioinformatics 28(21), 2819–2823 (2012)
30. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine learning 65(1), 31–78 (2006)
31. Willcutt, E.G., Pennington, B.F., DeFries, J.C.: Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. Journal of Abnormal Child Psychology 28(2), 149–159 (2000)
32. Yang, R., Berger, J.O.: Estimation of a covariance matrix using the reference prior. The Annals of Statistics pp. 1195–1211 (1994)