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

Ruifei Cui, Perry Groot, and Tom Heskes

{R.Cui,Perry.Groot,T.Heskes}@science.ru.nl

Institute for Computing and Information Sciences
Radboud University Nijmegen
Netherlands

19th November 2017

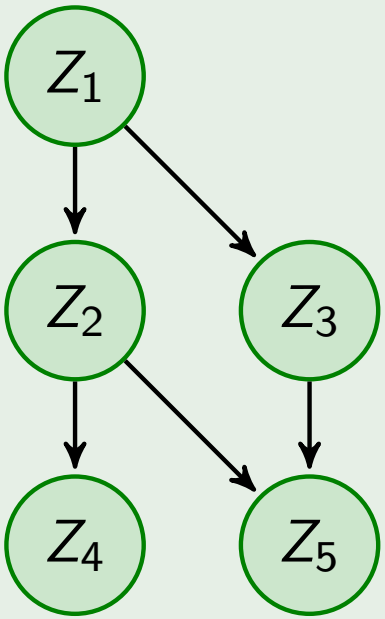
Outline

Outline

Causal Structure Learning (or Causal Discovery)

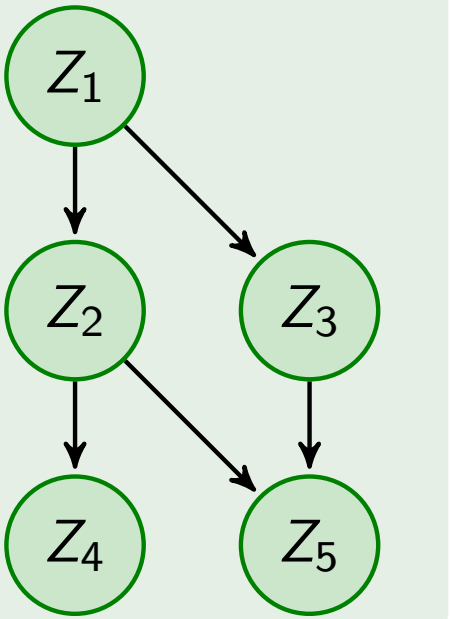
Causal Structure Learning (or Causal Discovery)

Example



Causal Structure Learning (or Causal Discovery)

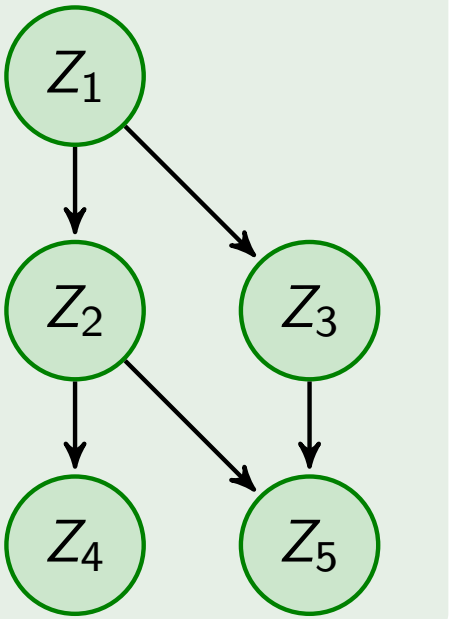
Example



- The vertices $\{Z_1, \ldots, Z_5\}$ denote random variables, and the directed edges represent causal relations between variables.

Causal Structure Learning (or Causal Discovery)

Example

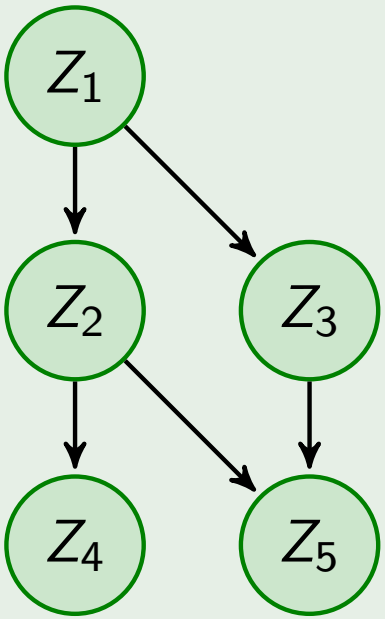


- The vertices $\{Z_1, \ldots, Z_5\}$ denote random variables, and the directed edges represent causal relations between variables.
- Given the graph, the data are assumed to be generated via

$$
\begin{aligned}
Z_1 &= \epsilon_1; \\
Z_2 &= aZ_1 + \epsilon_2; \\
Z_3 &= bZ_1 + \epsilon_3; \\
Z_4 &= cZ_2 + \epsilon_4; \\
Z_5 &= dZ_2 + eZ_3 + \epsilon_5.
\end{aligned}
$$

Causal Structure Learning (or Causal Discovery)

Example



- The vertices $\{Z_1, \ldots, Z_5\}$ denote random variables, and the directed edges represent causal relations between variables.
- Given the graph, the data are assumed to be generated via

$$
\begin{aligned}
Z_1 &= \epsilon_1; \\
Z_2 &= aZ_1 + \epsilon_2; \\
Z_3 &= bZ_1 + \epsilon_3; \\
Z_4 &= cZ_2 + \epsilon_4; \\
Z_5 &= dZ_2 + eZ_3 + \epsilon_5.
\end{aligned}
$$

Causal discovery aims to infer the (invariant parts of the) underlying graph from observational data.

The PC Algorithm

The PC algorithm is a reference causal discovery algorithm.

The PC Algorithm

The PC algorithm is a reference causal discovery algorithm.

The basic procedure

1. start with a fully connected undirected graph;
2. remove edges according to conditional independence tests, resulting in an undirected graph, called the skeleton;
3. apply various edge orientation rules, resulting in a Complete Partially Directed Acyclic Graph (CPDAG) that represents the invariant parts of the underlying graph.

The PC Algorithm

The PC algorithm is a reference causal discovery algorithm.

The basic procedure

1. start with a fully connected undirected graph;
2. remove edges according to conditional independence tests, resulting in an undirected graph, called the skeleton;
3. apply various edge orientation rules, resulting in a Complete Partially Directed Acyclic Graph (CPDAG) that represents the invariant parts of the underlying graph.

All work of the procedure is based on the conditional independence tests, therefore the focus of our paper.

The PC Algorithm for Gaussian Models

Conditional independence tests for Gaussian models

- When $\boldsymbol{Z} \sim \mathcal{N}(0, C)$, the PC algorithm considers the so-called partial correlation, denoted by $\rho_{uv|S}$, which can be estimated through the correlation matrix C.
- Using the classical decision theory with significance level α, this test boils down to
  $Z_u \perp\!\!\!\perp Z_v | \boldsymbol{Z}_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2)$,
  where $u \neq v$, $S \subseteq \{1, \ldots, p\} \backslash \{u, v\}$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian.

The PC Algorithm for Gaussian Models

Conditional independence tests for Gaussian models

- When $\boldsymbol{Z} \sim \mathcal{N}(0, C)$, the PC algorithm considers the so-called partial correlation, denoted by $\rho_{uv|S}$, which can be estimated through the correlation matrix C.
- Using the classical decision theory with significance level α, this test boils down to
  $Z_u \perp\!\!\!\perp Z_v | \boldsymbol{Z}_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2)$,
  where $u \neq v$, $S \subseteq \{1, \ldots, p\} \backslash \{u, v\}$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian.

Hence, the PC algorithm requires the sample correlation matrix $\hat{C}$ (to estimate $\rho_{uv|S}$) and the sample size n as input.

The PC Algorithm for Gaussian Copula Models

The PC Algorithm for Gaussian Copula Models

Definition (Gaussian Copula Model)

Consider two random vectors $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$, satisfying

1. $\boldsymbol{Z} \sim \mathcal{N}(0, C)$ (latent),
2. $Y_j = F_j^{-1}[\Phi(Z_j)] \; \forall \, j$ (observed),

where F_j^{-1} 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.

The PC Algorithm for Gaussian Copula Models

Definition (Gaussian Copula Model)

Consider two random vectors $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$, satisfying

1. $\boldsymbol{Z} \sim \mathcal{N}(0, C)$ (latent),
2. $Y_j = F_j^{-1}[\Phi(Z_j)] \; \forall \, j$ (observed),

where F_j^{-1} 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.

- In fully continuous cases, Harris and Drton (2013) proposed the Rank PC (RPC) algorithm.
- In mixed discrete and continuous cases, Cui et al. (2016) proposed the Copula PC (CoPC) algorithm.

The PC Algorithm for Gaussian Copula Models

Definition (Gaussian Copula Model)

Consider two random vectors $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$, satisfying

1. $\boldsymbol{Z} \sim \mathcal{N}(0, C)$ (latent),
2. $Y_j = F_j^{-1}[\Phi(Z_j)] \; \forall \, j$ (observed),

where F_j^{-1} 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.

However, both algorithms were proposed for complete data. In this paper, we aim to generalize them to data with missing values.

Two Missingness Mechanisms

Two Missingness Mechanisms

- When the missingness does not depend on the observed values, the data are said to be Missing Completely at Random (MCAR).

Two Missingness Mechanisms

- When the missingness does not depend on the observed values, the data are said to be Missing Completely at Random (MCAR).
- When the missingness depends on the observed values, the data are said to be Missing at Random (MAR).

Two Missingness Mechanisms

- When the missingness does not depend on the observed values, the data are said to be Missing Completely at Random (MCAR).
- When the missingness depends on the observed values, the data are said to be Missing at Random (MAR).
  For example, all people in a group are required to take a blood pressure test at time 1, while only those whose values at time 1 lie in the abnormal range need to take the test at time 2.

Two Missingness Mechanisms

- When the missingness does not depend on the observed values, the data are said to be Missing Completely at Random (MCAR).
- When the missingness depends on the observed values, the data are said to be Missing at Random (MAR).
  For example, all people in a group are required to take a blood pressure test at time 1, while only those whose values at time 1 lie in the abnormal range need to take the test at time 2.
- MCAR is a special case of MAR.

Outline

Rank PC Algorithm for Incomplete Data

Rank PC Algorithm for Incomplete Data

Provided that we have data $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ with missing values.

Rank PC Algorithm for Incomplete Data

Provided that we have data $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ with missing values.

Estimate the underlying correlation matrix

We use the so-called pairwise deletion strategy to handle with missing values.

- test the rank correlation between Y_j and Y_k based on complete observations for both variables, denoted by $\hat{C}_{jk}$;
- repeat this step over all pairs to get the underlying correlation matrix $\hat{C} = (\hat{C}_{jk})$.

Rank PC Algorithm for Incomplete Data

Effective sample size

Rank PC Algorithm for Incomplete Data

Effective sample size

- For complete data, the sample size n is needed in the conditional independence test,
  $Z_u \perp\!\!\!\perp Z_v | \boldsymbol{Z}_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2).$

Rank PC Algorithm for Incomplete Data

Effective sample size

- For complete data, the sample size n is needed in the conditional independence test,
  $Z_u \perp\!\!\!\perp Z_v | \boldsymbol{Z}_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2).$
- However, when the data contain missing values, the estimated correlations are less reliable than those estimated on complete data.

Rank PC Algorithm for Incomplete Data

Effective sample size

- For complete data, the sample size n is needed in the conditional independence test,
  $Z_u \perp\!\!\!\perp Z_v | \boldsymbol{Z}_S \Leftrightarrow \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2).$
- However, when the data contain missing values, the estimated correlations are less reliable than those estimated on complete data.
- To this end, we propose to replace the sample size (SS) with an effective sample size (ESS) in this test, acting as if the estimated correlations on incomplete data are in fact estimated from a smaller size of equivalent complete data.

Rank PC Algorithm for Incomplete Data

Estimate the effective sample size

Rank PC Algorithm for Incomplete Data

Estimate the effective sample size

- We first get the number of pairwise complete observations for variables Y_j and Y_k from data, denoted by $\hat{n}_{jk}$.
- Then, we consider two schemes when translating $\hat{n}_{jk}$ into the effective sample size to be used in the conditional independence tests.

Rank PC Algorithm for Incomplete Data

Estimate the effective sample size

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

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

This estimator is called the global ESS (GESS).

Rank PC Algorithm for Incomplete Data

Estimate the effective sample size

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

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

This estimator is called the global ESS (GESS).

2. We give a different ESS to different tests, since each test relies on a local structure involving only part of the variables, i.e.,

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

where $\hat{n}_{uv|S}$ is called the local ESS (LESS), defined as

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

Rank PC Algorithm for Incomplete Data

Call the standard PC algorithm

In the last step, we plug the estimated correlation matrix and the global (or local) effective sample size into the standard PC algorithm for causal discovery.

Copula PC Algorithm for Incomplete Data

Copula PC Algorithm for Incomplete Data

Review of the Copula PC algorithm for complete data

Given mixed data $\boldsymbol{Y}$ generated via a Gaussian copula model.

1. apply a Gibbs sampler to draw correlation matrix samples from the posterior distribution $P(C|\boldsymbol{Y})$;
2. use these samples to estimate the underlying correlation matrix and the effective sample size;
3. call the standard PC algorithm for causal discovery.

Copula PC Algorithm for Incomplete Data

Review of the Copula PC algorithm for complete data

Given mixed data $\boldsymbol{Y}$ generated via a Gaussian copula model.

1. apply a Gibbs sampler to draw correlation matrix samples from the posterior distribution $P(C|\boldsymbol{Y})$;
2. use these samples to estimate the underlying correlation matrix and the effective sample size;
3. call the standard PC algorithm for causal discovery.

Problem for incomplete data

When the data contain missing values, the Gibbs sampler in Step 1 does not work any longer. Therefore, the idea is to generalize the sampler to incomplete data, and then follow Step 2 and 3 for causal discovery.

Copula PC Algorithm for Incomplete Data

Gibbs sampler for mixed complete data

Given mixed complete data $\boldsymbol{Y}$, we obtain a sampling space $D(\boldsymbol{Y})$ for the latent pseudo Gaussian data $\boldsymbol{Z}$. Then, inference for the correlation matrix C proceeds by iterating the following two steps:

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

Copula PC Algorithm for Incomplete Data

Gibbs sampler for mixed complete data

Given mixed complete data $\boldsymbol{Y}$, we obtain a sampling space $D(\boldsymbol{Y})$ for the latent pseudo Gaussian data $\boldsymbol{Z}$. Then, inference for the correlation matrix C proceeds by iterating the following two steps:

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

Gibbs sampler for Gaussian incomplete data

Copula PC Algorithm for Incomplete Data

Gibbs sampler for mixed complete data

Given mixed complete data $\boldsymbol{Y}$, we obtain a sampling space $D(\boldsymbol{Y})$ for the latent pseudo Gaussian data $\boldsymbol{Z}$. Then, inference for the correlation matrix C proceeds by iterating the following two steps:

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

Gibbs sampler for Gaussian incomplete data

Given Gaussian incomplete data $\boldsymbol{Z} = (\boldsymbol{Z}_{obs}, \boldsymbol{Z}_{miss})$, we iterate the following two steps to impute missing values (step 1) and draw correlation matrix samples from the posterior (step 2):

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

Copula PC Algorithm for Incomplete Data

Gibbs sampler for mixed complete data

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

Gibbs sampler for Gaussian incomplete data

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

Gibbs sampler for mixed incomplete data

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

Outline

Results – Metrics

Results – Metrics

- The true positive rate (TPR) and the false positive rate (FPR) are used to evaluate the estimated skeleton.

Results – Metrics

- The true positive rate (TPR) and the false positive rate (FPR) are used to evaluate the estimated skeleton.
- The structural Hamming distance (SHD), counting the number of edge insertions, deletions, and flips in order to transfer the estimated CPDAG into the correct CPDAG, evaluates the CPDAG.

Results – Metrics

- The true positive rate (TPR) and the false positive rate (FPR) are used to evaluate the estimated skeleton.
- The structural Hamming distance (SHD), counting the number of edge insertions, deletions, and flips in order to transfer the estimated CPDAG into the correct CPDAG, evaluates the CPDAG.
- A higher TPR, a lower FPR, and a smaller SHD imply better performance.

Results

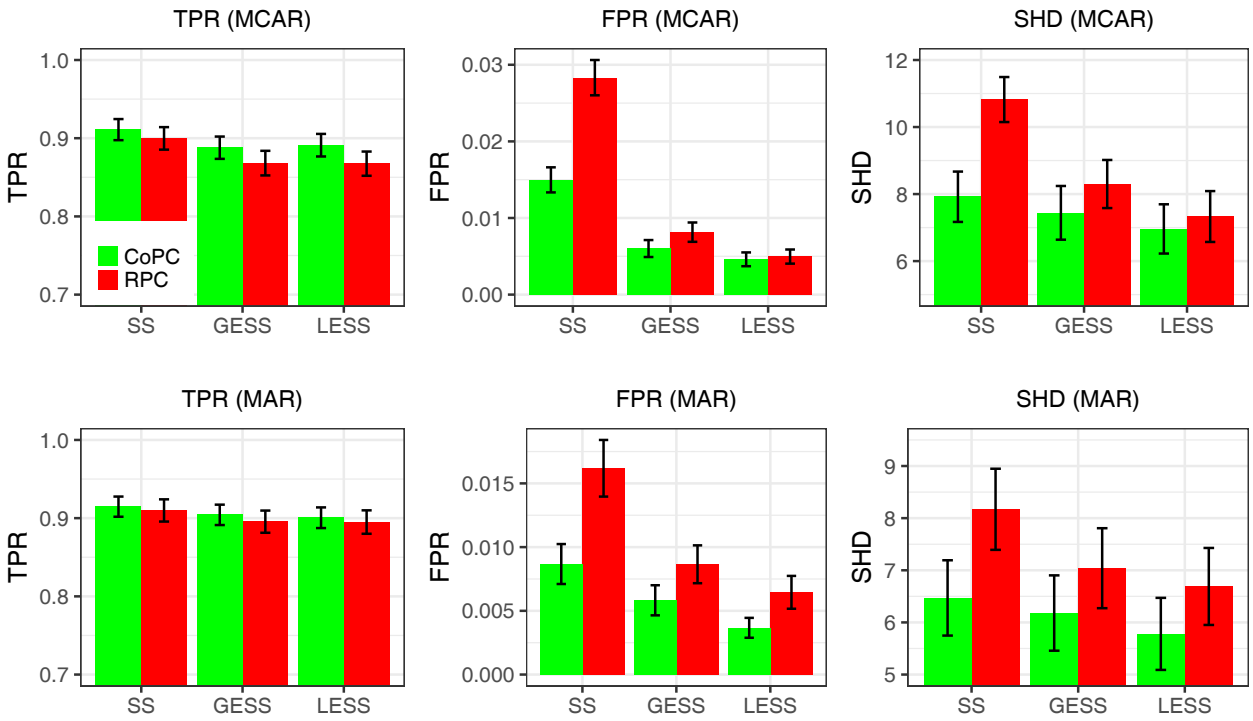
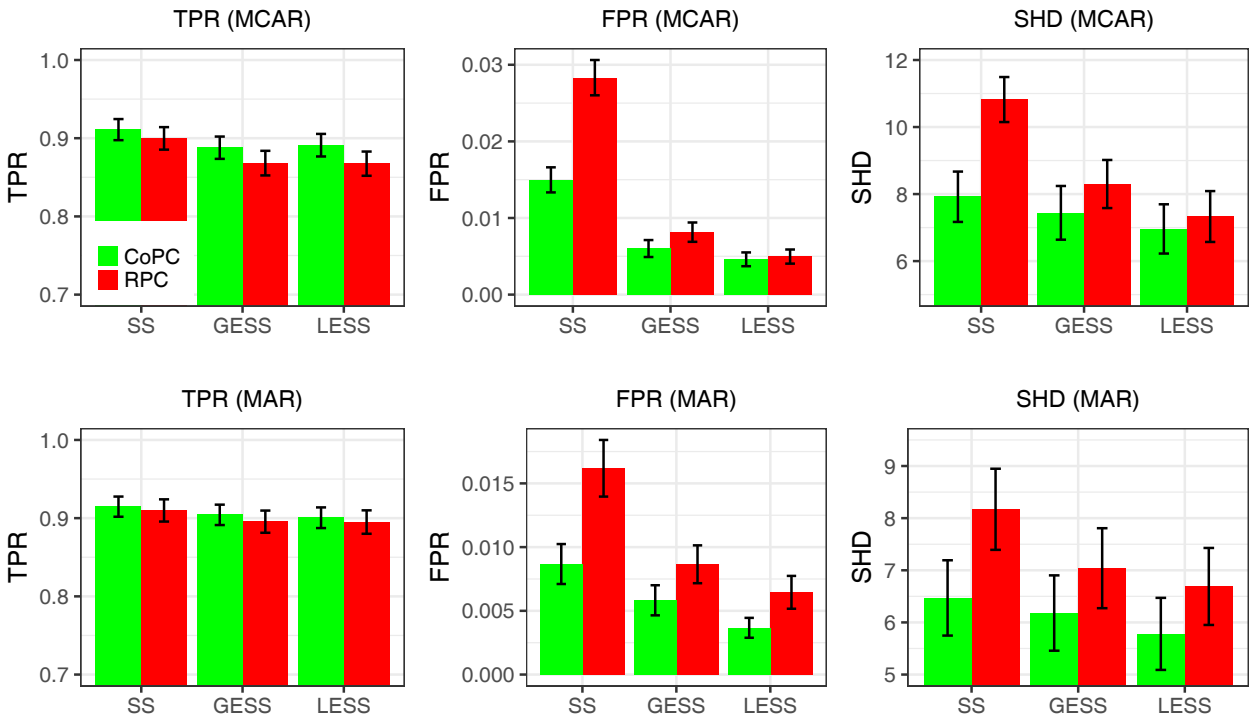


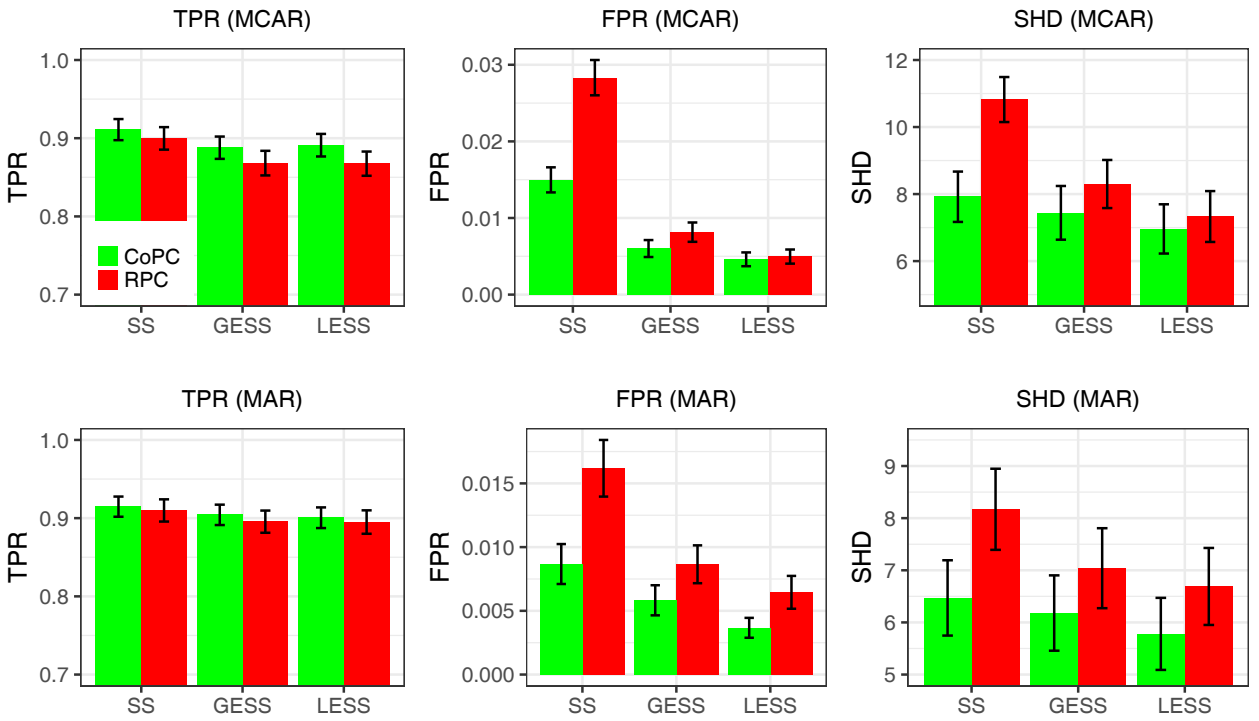
Figure: Performance of Rank PC (RPC) and Copula PC (CoPC) using SS, GESS, and LESS on incomplete data, showing the mean of TPR, FPR, and SHD over 100 experiments with 95% confidence interval under MCAR and MAR, respectively.

Results – Conclusions



- The usage of the effective sample size significantly improves the performance of Rank PC and Copula PC.

Results – Conclusions



- The usage of the effective sample size significantly improves the performance of Rank PC and Copula PC.
- Copula PC estimates a more accurate causal structure than Rank PC under MCAR and, even more so, under MAR.

Outline

Summary

Summary

- We propose two novel methods for causal discovery with missing data: Rank PC and Copula PC.
- They are provably correct for data under MCAR and MAR respectively.
- The two methods yield significant performance improvement over simpler approaches for missing data.

Thanks for your attention!