Learning the Causal Structure of Copula Models with Latent Variables

Ruifei Cui
Data Science
Radboud University
r.cui@science.ru.nl

Perry Groot
Data Science
Radboud University
perry.groot@cs.ru.nl

Moritz Schauer
Mathematical Institute
University of Leiden
m.r.schauer@math.leidenuniv.nl

Tom Heskes
Data Science
Radboud University
t.heskes@science.ru.nl

Abstract

A common goal in psychometrics, sociology, and econometrics is to uncover causal relations among latent variables representing hypothetical constructs that cannot be measured directly, such as attitude, intelligence, and motivation. Through measurement models, these constructs are typically linked to measurable indicators, e.g., responses to questionnaire items. This paper addresses the problem of causal structure learning among such latent variables and other observed variables. We propose the 'Copula Factor PC' algorithm as a novel two-step approach. It first draws samples of the underlying correlation matrix in a Gaussian copula factor model via a Gibbs sampler on rank-based data. These are then translated into an average correlation matrix and an effective sample size, which are taken as input to the standard PC algorithm for causal discovery in the second step. We prove the consistency of our 'Copula Factor PC' algorithm, and demonstrate that it outperforms the PC-MIMBuild algorithm and a greedy step-wise approach. We illustrate our method on a real-world data set about children with Attention Deficit Hyperactivity Disorder.

1 INTRODUCTION

Social scientists, psychologists, and many other scientists are usually interested in learning causal relations between latent variables that cannot be measured directly, e.g., attitude, intelligence, and motivation (see [15, 24], and Chapter 10 of [27] for more details). In order to get a grip on these latent concepts, one commonly-used strategy is to construct a measurement model for such a latent variable, in the sense that domain experts design a set of measurable "items" or survey "questions" that are considered to be indicators of the latent variable. For instance, in the study of Attention Deficit Hyperactivity Disorder (ADHD), 18 questions are designed to measure three latent variables: inattention, hyperactivity, and impulsivity [29]. In some other cases where it is difficult to design a measurement model due to the absence of domain knowledge or for other reasons, there are some off-the-shelf algorithms, e.g., BPC [24] and FOFC [15], for learning the measurement models from indicator data. In this paper, we focus on inferring the causal structure among latent variables, assuming that the measurement models are given. We also allow interactions between these latent variables and other (explicit) variables, e.g., subject characteristics like gender and age. Another issue we consider is that there are diverse types of variables in most real-world data: the questionnaire data in a survey is typically ordinal, whereas other variables might be binary, or continuous.

In this paper, we use a Gaussian copula factor model (the formal definition is given in Section 3) to describe such situations, in which a factor can be connected to either one or more observed variables (indicators). Factors with multiple indicators are used to model latent variables corresponding to psychological traits, such as attitude and intelligence. The copula model provides a good way of analyzing diverse types of variables, where the associations between variables are parameterized separately from their marginal distributions [13].

We propose the 'Copula Factor PC' algorithm for estimating the causal structure among factors of a Gaussian copula factor model, which is based on a two-step approach. The first step draws samples of the underlying correlation matrix, where the Gibbs sampler by [13] for Gaussian copula models is extended to Gaussian copula factor models by replacing the Wishart prior with the G-Wishart prior and adding a new strategy to sample latent factors. These samples are then translated into an aver-

age correlation matrix, and an effective sample size that is used to account for information loss incurred by discrete variables [9]. The second step takes the estimated correlation matrix and effective sample size as input to the standard PC algorithm [27] for causal discovery.

The rest of this paper is organized as follows. Section 2 reviews necessary knowledge and related work. Section 3 gives the definition of a Gaussian copula factor model. Section 4 describes our 'Copula Factor PC' algorithm, and introduces two alternative approaches: the PC-MIMBuild algorithm [24] and a greedy step-wise approach. Section 5 compares the 'Copula Factor PC' algorithm with the two alternative approaches on simulated data, and Section 6 gives an illustration on real-world data of ADHD patients. Section 7 concludes this paper and gives some discussion.

2 BACKGROUND

Causal discovery A graphical model is a graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, where the vertices $\boldsymbol{V} = \{X_1, \ldots, X_d\}$ correspond to random variables and the edges $\boldsymbol{E}$ 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 $X_i \to X_j$, X_i is called a parent of X_j. A distribution over a random vector $\boldsymbol{X}$ with $X_i \in \boldsymbol{V}$ is said to be Markov w.r.t. a DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, if $\boldsymbol{X}$ satisfies the *Causal Markov Condition*: each variable in the DAG $\mathcal{G}$ is independent of its non-descendants given its parents, which is also implied by the so-called *d-separation* [20]. A distribution is *faithful* w.r.t. a DAG $\mathcal{G}$ if there are no conditional independencies in the distribution that are not encoded by the Causal Markov Condition. If a distribution is both Markov and faithful w.r.t. a DAG $\mathcal{G}$, the DAG is called a *perfect map* of the distribution.

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). Arcs in a CPDAG suggest 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 implies 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 underlying DAG from observations.

The PC algorithm The PC algorithm [27], a reference algorithm for causal discovery, consists of two stages: adjacency search and orientation. The adjacency search starts with a fully connected undirected graph, and then recursively removes the edges according to conditional independence decisions, yielding the skeleton and separation sets. In the orientation stage, we first orient the unshielded triples according to the separation sets, and then orient as many of the remaining undirected edges as possible by applying the orientation rules repeatedly.

A key part of the procedure is to test for conditional independencies. 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 obtained by the correlation matrix C [1]. 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 $u \neq v$, $S \subseteq \{1, \ldots, d\} \backslash \{u, v\}$ and Φ is the cumulative distribution function of the standard Gaussian. Hence, the PC algorithm requires the correlation matrix C (to compute partial correlations $\rho_{uv|S}$) and the sample size n as input. Uniform consistency of the PC algorithm for Gaussian data is shown under some relatively mild assumptions on the sparsity of the true underlying structure [14] .

Harris & Drton [11] use rank correlations, typically Spearman's ρ and Kendall's τ, to replace the Pearson correlation, which extends the PC algorithm to the so-called nonparanormal models. The resulting 'Rank PC' algorithm performs as well as the PC algorithm using Pearson correlations on Gaussian data, yet much better on nonparanormal data. The PC algorithms using both Pearson and rank correlations require all univariate marginal distributions to be continuous. Cui et al. [9] extend the PC algorithm to mixed discrete and continuous data assumed to be drawn from a Gaussian copula model, where each observed variable is assumed to be induced by a latent Gaussian variable and the dependence between observed variables is determined by the correlation matrix of the latent variables. The resulting 'Copula PC' algorithm works well for mixed data, but requires each latent variable to have only a single indicator. Silva et at. [24] propose the PC-MIMBuild algorithm, which allows a latent variable to have multiple indicators, but it is limited to continuous observations and assumes that each latent variable has at least two indicators.

In this paper, we aim to generalize the PC algorithm to handle latent variables having one or more indicators and observations being either discrete or continuous.

3 GAUSSIAN COPULA FACTOR MODEL

Definition 1 (Gaussian Copula Factor Model). Consider a latent random (factor) vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$, a response random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^T$ and an observed random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T$, satisfying

$$\boldsymbol{\eta} \sim \mathcal{N}(0, C), \tag{2}$$

$$\boldsymbol{Z} = \Lambda\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3}$$

$$Y_j = F_j^{-1}\big(\Phi\big[Z_j/\sigma(Z_j)\big]\big), \forall j = 1, \ldots, p, \tag{4}$$

with $\Lambda = (\lambda_{ij})$ a $p \times k$ matrix of factor loadings ($k \leq p$), $\boldsymbol{\epsilon} \sim \mathcal{N}(0, D)$ Gaussian noise with $D = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, $\sigma(Z_j)$ the standard deviation of Z_j, and ${F_j}^{-1}(t) = \inf\{x : F_j(x) \geq t\}$ the pseudo-inverse of a cumulative distribution function F_j. This model is called a *Gaussian Copula Factor Model* with correlation matrix C, factor loadings Λ, and univariate margins F_j.

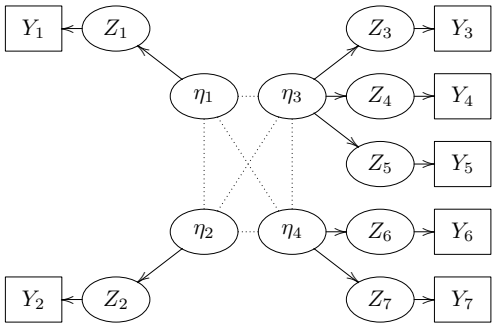


Figure 1: Gaussian copula factor model.

The model is also defined in [18], but the authors restrict the factors to be independent of each other while we allow for their interactions. An example of the model is shown in Figure 1. Our model is a combination of a Gaussian factor model (from $\boldsymbol{\eta}$ to $\boldsymbol{Z}$) and a Gaussian copula model (from $\boldsymbol{Z}$ to $\boldsymbol{Y}$). In the special case of a factor having a single response (thus a single observed variable), e.g., $\eta_1 \to Z_1 \to Y_1$, it reduces to a Gaussian copula model where we set $\lambda_{11} = 1$ and $\epsilon_1 = 0$, thus $Y_1 = F_1^{-1}(\Phi[\eta_1])$.

In the typical design for questionnaires, one tries to get a grip on a latent concept through a particular set of well-designed questions [16, 4], which implies that a factor (latent concept) in our model is connected to multiple indicators (questions) while an indicator is only used to measure a single factor, as shown in Figure 1. This kind of measurement model is called a pure measurement model (Definition 2 of [23]). Throughout this paper, we assume that all measurement models are given and pure, which makes that there is only a single non-zero entry in each row of the factor loadings matrix Λ. This inductive bias about the sparsity pattern of Λ is fully motivated by the typical design of a measurement model.

In what follows, we transform the Gaussian copula factor model into an equivalent model, which we will use for inference in the next section. We consider an integrated random vector $\boldsymbol{X} = (\boldsymbol{Z}^T, \boldsymbol{\eta}^T)^T$, which is still multivariate Gaussian, and obtain its covariance matrix

$$\Sigma = \begin{bmatrix} \Lambda C \Lambda^T + D & \Lambda C \\ C\Lambda^T & C \end{bmatrix}, \tag{5}$$

and precision matrix

$$\Omega = \Sigma^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}\Lambda \\ -\Lambda^T D^{-1} & C^{-1} + \Lambda^T D^{-1} \Lambda \end{bmatrix}. \tag{6}$$

Since D is diagonal and Λ only has one non-zero entry per row, Ω contains many intrinsic zeros. The sparsity pattern of such $\Omega = (\omega_{ij})$ can be represented by an undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where $(i, j) \notin \boldsymbol{E}$ whenever $\omega_{ij} = 0$ by construction. Then, a Gaussian copula factor model can be transformed into an equivalent model controlled by a single precision matrix Ω, which in turn is constrained by G, i.e., $P(\boldsymbol{X}|C, \Lambda, D) = P(\boldsymbol{X}|\Omega_G)$.

Definition 2 (G-Wishart Distribution [22]). Given an undirected graph $G = (\boldsymbol{V}, \boldsymbol{E})$, a zero-constrained random matrix Ω has a G-Wishart distribution, if its density is

$$p(\Omega|G) = \frac{|\Omega|^{(\nu-2)/2}}{I_G(\nu, \Psi)} \exp\left[-\frac{1}{2}\operatorname{tr}(\Psi\Omega)\right] \mathbb{1}_{\Omega \in M^+(G)},$$

with $M^+(G)$ the space of symmetric positive definite matrices with off-diagonal elements $\omega_{ij} = 0$ whenever $(i, j) \notin \boldsymbol{E}$, ν the number of degrees of freedom, Ψ a scale matrix, $I_G(\nu, \Psi)$ the normalizing constant, and $\mathbb{1}$ the indicator function.

The G-Wishart distribution is the conjugate prior of precision matrices Ω that are constrained by a graph G [22]. That is, given the G-Wishart prior, i.e., $P(\Omega|G) = \mathcal{W}_G(\nu_0, \Psi_0)$ and data $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})^T$ drawn from $\mathcal{N}(0, \Omega^{-1})$, the posterior for Ω is another G-Wishart distribution:

$$P(\Omega|G, \boldsymbol{X}) = \mathcal{W}_G(\nu_0 + n, \Psi_0 + \boldsymbol{X}^T\boldsymbol{X}).$$

When the graph G is fully connected, the G-Wishart distribution reduces to a Wishart distribution [17]. Placing a G-Wishart prior on Ω is equivalent to placing an inverse-Wishart on C, a product of multivariate normals on Λ, and an inverse-gamma on the diagonal elements of D. With a diagonal scale matrix Ψ_0 and the number of degrees of freedom ν_0 equal to the number of factors plus one, the implied marginal densities between any pair of factors are uniformly distributed in the interval $[-1, 1]$ [3].

4 METHODS

In this section, we propose a Bayesian inference method for Gaussian copula factor models, based on which we derive our 'Copula Factor PC' algorithm. Then, we introduce two alternative approaches.

4.1 INFERENCE FOR GAUSSIAN COPULA FACTOR MODEL

For a Gaussian copula model, Hoff [13] proposed a likelihood that only concerns the ranks among observations, which is derived as follows. Since the transformation $Y_j = F_j^{-1}\big(\Phi\big[Z_j\big]\big)$ is non-decreasing, observing $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n,j})^T$ implies a partial ordering on $\boldsymbol{z}_j = (z_{1,j}, \ldots, z_{n,j})^T$, namely, $\boldsymbol{z}_j$ must lie in the space restricted by $\boldsymbol{y}_j$:

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

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

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

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

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

where S is the correlation matrix over $\boldsymbol{Z}$.

Following the same argumentation, the likelihood in our Gaussian copula factor model reads

$$P(\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y})|\boldsymbol{\eta}, \Omega, F_1, \ldots, F_p) = P(\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y})|\boldsymbol{\eta}, \Omega),$$

which is independent of the margins F_j.

For the Gaussian copula factor model, inference for the precision matrix Ω of the vector $\boldsymbol{X} = (\boldsymbol{Z}^T, \boldsymbol{\eta}^T)^T$ can now proceed via construction of a Markov chain having its stationary distribution equal to $P(\boldsymbol{Z}, \boldsymbol{\eta}, \Omega|\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}), G)$, where we ignore the values for $\boldsymbol{\eta}$ and $\boldsymbol{Z}$ in our samples. The prior graph G is uniquely determined by the sparsity pattern of the loading matrix $\Lambda = (\lambda_{ij})$ and the residual matrix D (see Equation 6), which in turn is uniquely decided by the pure measurement models. The Markov chain can be constructed by iterating the following three steps:

1. **Sample** $\boldsymbol{Z}$: $\boldsymbol{Z} \sim P(\boldsymbol{Z}|\boldsymbol{\eta}, \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}), \Omega)$;
   Since each coordinate Z_j directly depends on only one factor, i.e., η_q such that $\lambda_{jq} \neq 0$, we can sample each of them independently through $Z_j \sim P(Z_j|\eta_q, \boldsymbol{z}_j \in \mathcal{D}(\boldsymbol{y}_j), \Omega)$.
2. **Sample** $\boldsymbol{\eta}$: $\boldsymbol{\eta} \sim P(\boldsymbol{\eta}|\boldsymbol{Z}, \Omega)$;
3. **Sample** Ω: $\Omega \sim P(\Omega|\boldsymbol{Z}, \boldsymbol{\eta}, G)$.

A Gibbs sampler that implements the Markov chain is summarized in Algorithm 1.

Algorithm 1 Gibbs sampler for Gaussian copula factor model

Require: Measurement models (decide sparsity of Λ and thus G), and indicator data $\boldsymbol{Y}$.

Step 1: sample $\boldsymbol{Z} \sim P(\boldsymbol{Z}|\boldsymbol{\eta}, \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}), \Omega)$.
for $j \in \{1, \ldots, p\}$ **do**
  $q =$ factor index of Z_j
  $a = \Sigma_{[j,q+p]}/\Sigma_{[q+p,q+p]}$
  $\sigma_j^2 = \Sigma_{[j,j]} - a\Sigma_{[q+p,j]}$
  for $y \in \text{unique}\{y_{1,j}, \ldots, y_{n,j}\}$ **do**
    $z_l = \max\{z_{i,j} : y_{i,j} < y\}$
    $z_u = \min\{z_{i,j} : y < y_{i,j}\}$
    for i such that $y_{i,j} = y$ **do**
      $\mu_{i,j} = \boldsymbol{\eta}_{[i,q]} \times a$
      $u_{i,j} \sim \mathcal{U}\big(\Phi\big[\frac{z_l - \mu_{i,j}}{\sigma_j}\big], \Phi\big[\frac{z_u - \mu_{i,j}}{\sigma_j}\big]\big)$
      $z_{i,j} = \mu_{i,j} + \sigma_j \times \Phi^{-1}(u_{i,j})$
    end for
  end for
end for
Step 2: sample $\boldsymbol{\eta} \sim P(\boldsymbol{\eta}|\boldsymbol{Z}, \Omega)$.
$A = \Sigma_{[\boldsymbol{\eta},\boldsymbol{Z}]}\Sigma_{[\boldsymbol{Z},\boldsymbol{Z}]}^{-1}$
$B = \Sigma_{[\boldsymbol{\eta},\boldsymbol{\eta}]} - A\Sigma_{[\boldsymbol{Z},\boldsymbol{\eta}]}$
for $i \in \{1, \ldots, n\}$ **do**
  $\boldsymbol{\mu}_i = (\boldsymbol{Z}_{[i,:]}A^T)^T$
  $\boldsymbol{\eta}_{[i,:]} \sim \mathcal{N}(\boldsymbol{\mu}_i, B)$
end for
$\boldsymbol{\eta}_{[:,j]} = \boldsymbol{\eta}_{[:,j]} \times \text{sign}(\text{Cov}[\boldsymbol{\eta}_{[:,j]}, \boldsymbol{Z}_{[:,f(j)]}]), \forall j$, where $f(j)$ is the index of the first indicator of η_j.
Step 3: sample $\Omega \sim P(\Omega|\boldsymbol{Z}, \boldsymbol{\eta}, G)$.
$\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{\eta})$
$\Omega \sim \mathcal{W}_G(\nu_0 + n, \Psi_0 + \boldsymbol{X}^T\boldsymbol{X})$
$\Sigma = \Omega^{-1}$
$\Sigma_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}, \forall i, j$

Identifiability of C: Without additional constraints, the correlation matrix C over factors is non-identifiable [2]. More precisely, given a decomposable covariance matrix $S = \Lambda C\Lambda^T + D$, we can always replace Λ with ΛU and C with $U^{-1}CU^{-T}$ to obtain an equivalent decomposition $S = (\Lambda U)(U^{-1}CU^{-T})(U^T\Lambda^T) + D$, where U is a $k \times k$ invertible matrix. Since Λ only has one non-zero entry per row in our model, U can only be diagonal to ensure that ΛU has the same sparsity pattern as Λ (see Lemma 3 in Supplement). Thus, from the same S, we get a class of solutions for C, i.e., $U^{-1}CU^{-1}$, where

U can be any invertible diagonal matrix. However, we find that all members in this class encode the same set of conditional independencies (see Lemma 4 in Supplement), and therefore imply the same causal structure [27]. Hence, any solution in this class is appropriate for finding the underlying causal structure among latent variables.

In order to get a unique solution for C, we impose two sufficient identifying conditions: 1) restrict C to be a correlation matrix; 2) force the first non-zero entry in each column of Λ to be positive (see Lemma 5 in Supplement). Condition 1 is implemented via line 28 in Algorithm 1. As for the second condition, we force the covariance between a factor and its first indicator to be positive (line 23), which is equivalent to Condition 2. One could also choose one's favorite constraints for identifying C, as long as the unique solution belongs to the class $U^{-1}CU^{-1}$.

4.2 COPULA FACTOR PC ALGORITHM

By iterating the steps in Algorithm 1 and extracting the submatrix over $\boldsymbol{\eta}$, we can draw samples of C, 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., $\hat{C} = \frac{1}{m}\sum_{i=1}^{m} C^{(i)}$. As for the effective sample size $\hat{n}$, we build upon the idea in [9], that is, taking the posterior distribution's degrees of freedom ν as an approximation to $\hat{n}$. Theorem 1 (the proof is provided in the Supplement) suggests a procedure to estimate the degrees of freedom of a G-Wishart distribution.

Theorem 1. *Consider a random matrix Ω following a G-Wishart distribution with graph $G = (\boldsymbol{V}, \boldsymbol{E})$ as well as parameters ν and Ψ, i.e., $\Omega \sim \mathcal{W}_G(\nu, \Psi)$. Let $\Sigma = \Omega^{-1}$ and $\widetilde{\Sigma}$ be the normalized matrix of Σ, i.e., $\widetilde{\Sigma}_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$. Then, for large ν, we have*

$$\mathrm{Var}\left[\widetilde{\Sigma}_{ij}\right] \approx \frac{(1 - (\mathrm{E}\left[\widetilde{\Sigma}_{ij}\right])^2)^2}{\nu}\,, \tag{7}$$

for off-diagonal elements $\widetilde{\Sigma}_{ij}$ whenever $(i, j) \in \boldsymbol{E}$.

From the theorem, we have that all off-diagonal elements of the latent correlation matrix satisfy Equation (7), because the prior subgraph over latent factors is fully connected. Therefore, we estimate $\hat{n}$ as follows

$$\hat{n} = \frac{1}{k(k-1)}\sum_{i \neq j} \nu_{ij}, \;\; \text{where} \;\; \nu_{ij} = \frac{(1 - (\mathrm{E}\left[C_{ij}\right])^2)^2}{\mathrm{Var}\left[C_{ij}\right]}.$$

The 'Copula Factor PC' (CFPC) algorithm arises when taking the estimated correlation matrix $\hat{C}$ and the effective sample size $\hat{n}$ (to replace the n in Equation 1) as the input to the standard PC algorithm.[1] The CFPC algorithm is consistent, as shown in Theorem 2 (see proof in the Supplement).

Theorem 2 (Consistency of the CFPC algorithm). *Let $\boldsymbol{Y}_n = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ be independent observations drawn from a Gaussian copula factor model. If 1) the measurement model per factor is known and pure; and 2) the distribution over factors is faithful to a DAG $\mathcal{G}$, then*

$$\lim_{n\to\infty} P\big(\hat{\mathcal{M}}_n(\mathcal{G}) = \mathcal{M}(\mathcal{G})\big) = 1\,,$$

where $\hat{\mathcal{M}}_n(\mathcal{G})$ is the output of the CFPC algorithm and $\mathcal{M}(\mathcal{G})$ is the Markov equivalent class of the true underlying DAG $\mathcal{G}$.

4.3 ALTERNATIVE APPROACHES

The PC-MIMBuild algorithm The original PC-MIMBuild algorithm only works for continuous data. Here, we extend it to mixed cases by learning the correlation matrix of response variables via the Gibbs sampler by [13] and taking it as input to the original PC-MIMBuild. We further generalize the PC-MIMBuild algorithm to handle latent factors with just a single indicator, by replacing the conditional independence testing method designed only for factors with at least two indicators (Theorem 19 in [24]) with a test based on partial correlation. See Supplement B for more details.

A greedy step-wise approach This approach first extracts the measurement model of a factor with multiple indicators, e.g., the subpart of Figure 1 consisting of the variables $\{\eta_3, Z_3, Z_4, Z_5, Y_3, Y_4, Y_5\}$. Then, it uses off-the-shelf techniques [10] to fit such a model and obtain pseudo-data of the factor (factor scores). Using pseudo-data for factors with multiple indicators together with real data for factors with a single indicator, the 'Copula PC' algorithm is next applied for causal discovery. We refer to this approach as the greedy step-wise PC algorithm, whose pseudo-code is written out step by step in the Supplement C. One disadvantage of this approach is that it can overestimate the effective sample size when treating the pseudo-data at the same footing as real data. This might incur many false positives, as we will indeed observe in the experiment section.

5 SIMULATION STUDY

In this section, we compare our 'Copula Factor PC' algorithm (CFPC) with the PC-MIMBuild algorithm (MBPC) and the greedy step-wise PC algorithm (GSPC)

[1] The R code is publicly available in `https://github.com/cuiruifei/CopulaFactorModel`.

on simulated data. Kalisch & Buhlmann [14] provide a procedure to generate random DAGs and simulate normally distributed samples that are faithful to them. It first generates a $k \times k$ adjacency matrix A representing a random DAG: 1) generate a $k \times k$ zero matrix, 2) randomly set entries in the lower-triangle area to be *one* with probability s (measuring the sparseness), 3) change the *ones* to be random weights in the interval $[0.1, 1]$. Given the adjacency matrix A, values of a random vector $\boldsymbol{\eta}$ are drawn recursively via

$$\eta_i = \sum_{k<i} A_{ik}\eta_k + \epsilon_i \,,$$

with each $\epsilon_i \sim \mathcal{N}(0, 1)$. Following this procedure, we simulate the factors of a Gaussian copula factor model, i.e., the $\boldsymbol{\eta}$ in Equation (2). Then, the edge weights from factors to response variables (non-zero elements of Λ in Equation 3) are uniformly drawn from the interval $[0.1, 1]$. We next generate response variables using Equation (3) together with standard Gaussian noise. After discretizing some response variables, we obtain data following a Gaussian copula factor distribution.

Three metrics are used to evaluate the algorithms: the true and false positive rate (TPR and FPR) for assessing the skeleton, and the structural Hamming distance (SHD), counting the number of edge insertions, deletions, and flips to transfer the estimated CPDAG into the correct CPDAG [28], for assessing the CPDAG. A higher TPR, a lower FPR, and a smaller SHD imply better performance. We set the significance level in the PC algorithm to $\alpha = 0.01$ (experiments with other values done suggest the same conclusion) and the sparseness parameter in generating DAGs to $s = 2/(k-1)$, such that the average neighbors of each node is 2 [14]. For the Gibbs sampler, the first 500 samples (burn-in) are discarded and the next 500 samples are stored. We test the algorithms for different numbers of factors $k \in \{4, 10\}$, and sample sizes $n \in \{500, 1000, 2000\}$.

Evaluation on Gaussian data We first consider the case where the observed data are Gaussian and all factors have multiple indicators, since this matches the assumptions of the original PC-MIMBuild algorithm. The number of indicators per factor is randomly chosen from 3 to 10, to mimic typical real-world datasets [25, 29].

Figure 2 shows the results, providing the mean of TPR, FPR, and SHD over 100 repeated experiments with errorbars representing 95% confidence intervals. First, we see that CFPC performs clearly better than MBPC w.r.t. TPR despite an indistinguishable performance w.r.t. FPR (CFPC is slightly better than MBPC for $k = 4$ while the other way around for $k = 10$). Therefore, w.r.t. the overall metric SHD, CFPC significantly outperforms MBPC. Our analysis is that MBPC tests for conditional independencies between all pairs of indicators and claims a dependence between factors even if just one of the pairs fails the test. This multiple testing approach, although elegant in theory, is difficult to make robust for largely varying numbers of indicators and sizes of the conditioning set. Second, while CFPC and GSPC report similar TPR scores, CFPC shows a clear advantage over GSPC w.r.t. FPR (thus a better SHD than GSPC), which becomes more prominent for a larger sample size. This is because the correlations between factors are estimated indirectly through their indicators, which makes the correlations less reliable than those estimated directly through the observed data. The effective sample size used in CFPC naturally incorporates the reduced reliability, whereas GSPC that still uses the original sample size rejects the null hypothesis of conditional independence more easily, resulting in more false positives.

Evaluation on mixed data We now focus on mixed data, in which two cases are considered: 1) all factors have multiple indicators; 2) half of the factors have multiple indicators and half only have a single indicator. When a factor has multiple indicators, the number of indicators per factor is randomly chosen from 3 to 10, and all such indicators are discretized into ordinal variables where the number of levels per variable is randomly chosen from 2 to 5. For factors with a single indicator, we discretize half into ordinal variables (from 2 to 5 levels) and keep the other half continuous.

Figures 3 and 4 summarize the experimental results, providing the mean of TPR, FPR, and SHD over 100 repeated experiments with 95% confidence intervals. From Figure 3a, we first see that GSPC is slightly better than CFPC w.r.t. TPR while GSPC and CFPC show a clear advantage over MBPC. Second, MBPC is rather sensitive to sample sizes in cases with only multiple indicators, where a small sample size incurs a poor performance. Figure 3b shows that CFPC is significantly better than GSPC w.r.t. FPR, which becomes more prominent in cases with only multiple indicators and larger sample sizes. This is because the effective sample size in CFPC better than GSPC represents the uncertainty in the partial correlation estimates and then incurs less false positives. CFPC also shows clear advantages over MBPC w.r.t. FPR when the number of factors is 4 ($k = 4$), whereas MBPC works slightly better than CFPC when $k = 10$. As for the overall metric SHD shown in Figure 4, CFPC and GSPC perform clearly better than MBPC in almost all situations because of the bad performance of MBPC w.r.t. TPR. Meanwhile, we can see that CFPC generates a more accurate CPDAG than GSPC, in particular for larger sample sizes. This is because our proposed infer-

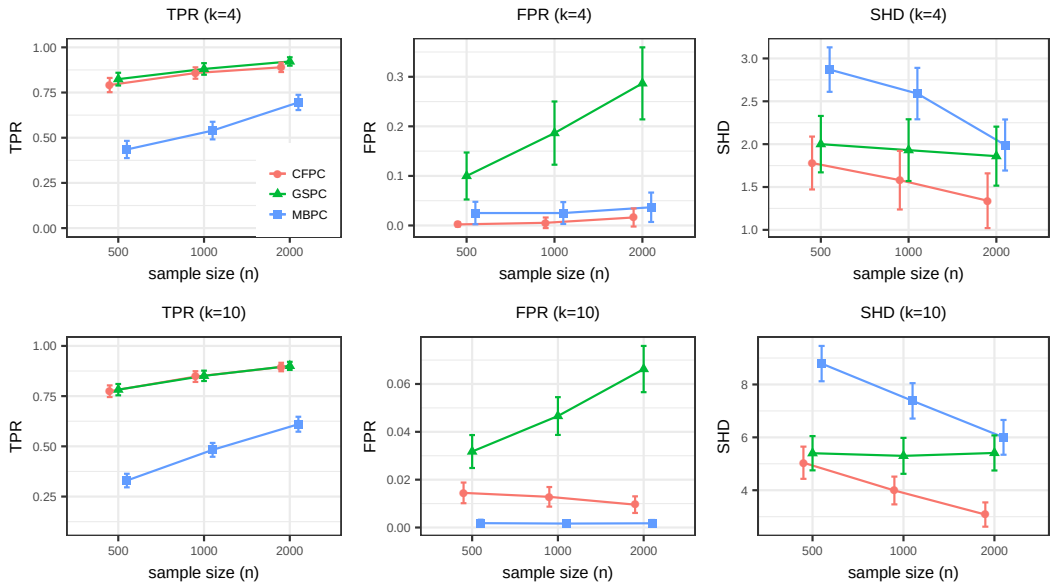


Figure 2: TPR, FPR, and SHD of CFPC, GSPC, and MBPC over different sample sizes when the data are fully Gaussian and all factors have multiple indicators, showing the mean over 100 experiments together with 95% confidence intervals. The two rows represent the results when the number of latent factors is 4 and 10 respectively.

ence procedure more accurately estimates the correlation matrix (not shown here) and, through the effective sample size, better represents the uncertainty in the correlation estimates than the greedy step-wise method. In a nutshell, our 'Copula Factor PC' algorithm, outperforms its two competitors in almost all situations.

6 REAL-WORLD APPLICATION

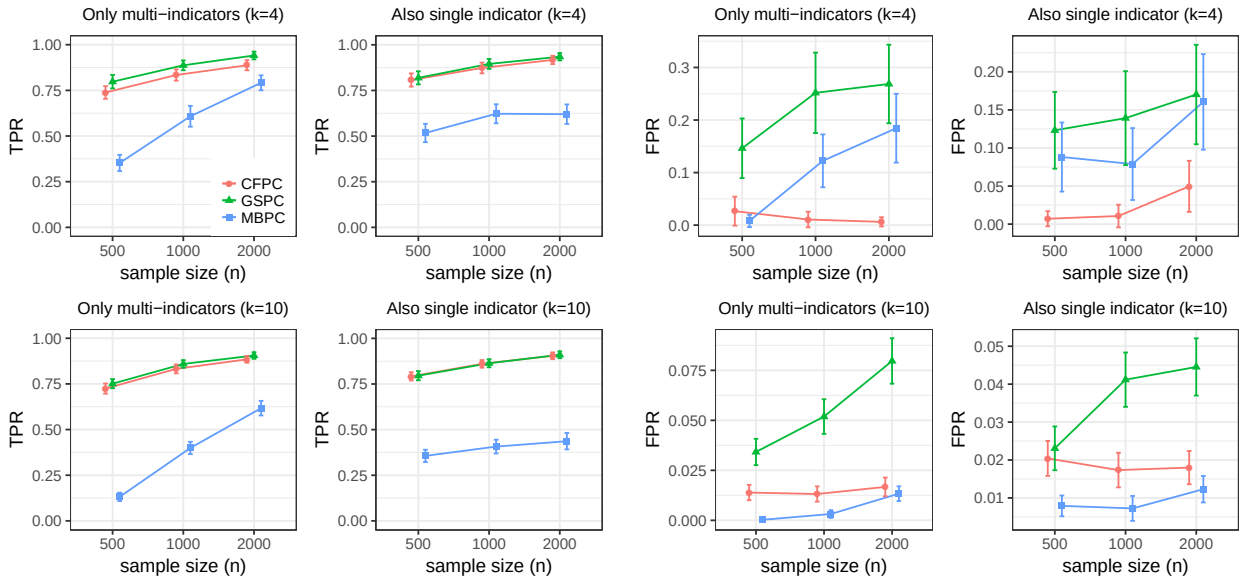
In this section, we give an illustration on a real-world dataset collected by [30] that includes 236 children with Attention Deficit Hyperactivity Disorder (ADHD) and 406 controls. We focus on 4 (explicit) variables that are related to ADHD symptoms: gender (Gen), Age, verbal IQ (VIQ), performance IQ (PIQ), as well as 18 questions that are designed to measure three latent concepts: inattention (Inatt), hyperactivity (Hyper), and impulsivity (Impul). The first 9 questions (Q1-Q9) are designed to measure 'Inatt', while the next 5 questions (Q10-Q14) and the last 4 questions (Q15-Q18) are used to measure 'Hyper' and 'Impul' respectively [29]. All the questions are ordinal with four levels: never (0), sometimes (1), frequently (2), and always (3).

Our task is to infer the causal structure among the 4 variables and 3 latent concepts from observations of the 4 variables and 18 questions. We run our 'Copula Factor PC' algorithm (using the order-independent version of the PC algorithm [7]) on this dataset and enforce the prior knowledge that no variables cause gender. The resulting graph is shown in Figure 5, in which double arrows '$\Rightarrow$' represent the mapping from the three latent concepts to their corresponding questions (known) and other edges are those learned by our algorithm.

First, in the inferred model, we find that 'Gen' has a direct causal influence on 'Inatt'. The finding is in the expected direction, namely males are at an increased risk of inattention, hyperactivity, and impulsivity problems. Meta-analyses in population-based samples suggested that males are 24 times more likely to meet full criteria for ADHD than females [31] and in clinically referred ADHD samples, the gender ratio was about 5:1 [19].

Second, the causal model implies that there is a significant causal path from inattention to hyperactivity (and subsequently to impulsivity), but not the other way around. It suggests that factors that cause inattention affect hyperactivity/impulsivity downstream of that, whereas those factors that lead to high hyperactivity/impulsivity do not necessarily lead to higher inattention. This causal path was previously observed in this sample and was also confirmed in two independent ADHD samples [26].

Third, the causal direction of the associations between verbal IQ and inattention as well as impulsivity is not clear from our model. Both interpretations seem reasonable. Previous studies suggest that ADHD is associated with lower (verbal) IQ, and particularly attention problems have been found to be strong predictors for lower



(a) TPR (b) FPR

Figure 3: (a) TPR of CFPC, GSPC, and MBPC for the case where all factors have multiple indicators (left column) and the case where half of the factors have multiple indicators while the other half have a single indicator (right column), showing the mean over 100 experiments together with 95% confidence intervals. The two rows represent the results when the number of latent factors is 4 and 10 respectively. (b) FPR for the same experiments as in (a).

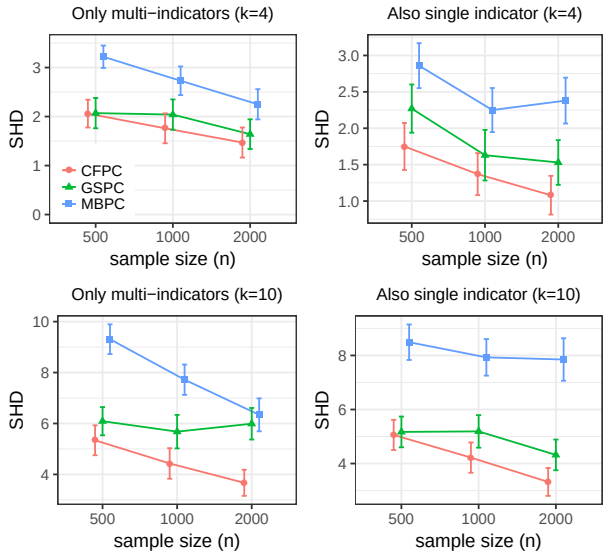


Figure 4: SHD of CFPC, GSPC, and MBPC, showing the mean over 100 experiments together with 95% confidence intervals, for the same experiments as in Figure 3.

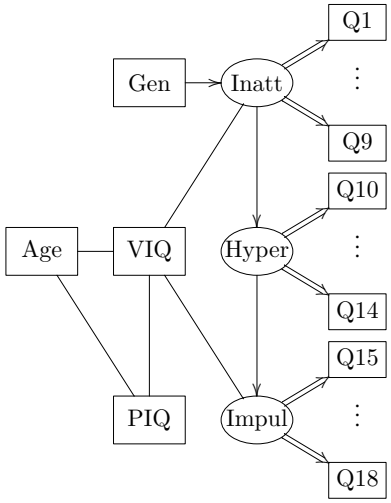


Figure 5: The resulting causal graph obtained by the 'Copula Factor PC' algorithm on the ADHD dataset, in which double arrows '⇒' represent the mapping from latent concepts to their corresponding questions (known) and other edges are those learned by our algorithm.

IQ and poorer academic performance [12].

To conclude, using the Copula Factor PC algorithm in an ADHD sample allows us to infer causal relations between the different ADHD traits and generic factors (age, gender, and IQ). This enhances knowledge of the causal structure of ADHD (e.g., by answering the question whether inattention is causing hyperactivity, or vice versa), which may have significant clinical implications, as it may inform therapeutic interventions.

7 CONCLUSION AND DISCUSSION

In this paper, we focused on learning causal relations between latent variables with pre-designed or pre-fitted measurement models. Our typical use case is that of psychological constructs that are linked to responses on questionnaire items. To the best of our knowledge, we are the first to propose a provably convergent algorithm that is able to recover the underlying causal structure between such factors and other observed variables, which can be both discrete and continuous.

In the experiments, our 'Copula Factor PC' algorithm clearly outperformed both the PC-MIMBuild algorithm and the greedy step-wise approach. PC-MIMBuild tests for conditional independencies between all pairs of indicators and concludes that the latent factors are dependent even if just one of the pairs fails the independence test. In our experience, this multiple testing approach, although elegant in theory, is difficult to make robust for largely varying numbers of indicators and sizes of the conditioning set. The 'Copula Factor PC' algorithm more naturally appears to find the right balance between true positives and false positives under varying conditions. It improves upon the greedy step-wise approach by estimating the full correlation matrix instead of individual sub-parts, which increases the power of the conditional independence tests.

Our approach extends earlier work, particularly [9] and [11], with various novel and essential ingredients needed to handle latent variables. Compared to [9], we replaced the Wishart prior with a G-Wishart distribution over factors and indicator variables, whose structure directly follows from the measurement model. The corresponding marginal prior on the factors is then still a Wishart distribution, which can be chosen such that the pairwise correlations are uniformly distributed. As in [11], but unlike [9], we can prove that our procedure is consistent. In the Supplement we show that, although the correlation matrix over factors itself is non-identifiable, all characteristics that relate to the identification of the correct causal structure can be consistently recovered.

While we considered the PC algorithm for inferring the underlying causal structure, one could plug in other standard algorithms like FCI [27], GES [5], or the recent improvements [6, 8, 32]. We further focused on so-called pure measurement models [24, 15], which is the major simplifying assumption of our procedure. We would argue that this is often satisfied, since it is the way in which questionnaires are typically designed by domain experts and that allows for a specific interpretation of the factors (e.g., a predefined set of items relates to the concept "hyperactivity", another non-overlapping set of items to the concept "inattention"). If the measurement models are not given, they can be learned using off-the-shelf algorithms, such as BPC [24] and FOFC [15], which output pure measurement models.

Acknowledgements

We thank Anoek Sluiter-Oerlemans for the profound analysis about experimental results on the ADHD dataset, and Ioan Gabriel Bucur for valuable discussions on the proof of Theorem 2. This research has been partially financed by the Netherlands Organisation for Scientific Research (NWO) under project 617.001.451.

References

[1] Anderson, Theodore Wilbur. *An introduction to multivariate statistical analysis*. John Wiley & Sons, 2003.

[2] Anderson, Theodore Wilbur and Rubin, Herman. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pp. 111–150, Berkeley, Calif., 1956. University of California Press.

[3] Barnard, John, McCulloch, Robert, and Meng, Xiao-Li. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pp. 1281–1311, 2000.

[4] Byrne, Barbara M. *Structural equation modeling with EQS: Basic concepts, applications, and programming*. Routledge, 2013.

[5] Chickering, David Maxwell. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[6] Claassen, Tom and Heskes, Tom. A Bayesian approach to constraint based causal inference. In *UAI*, pp. 207–216, 2012.

[7] Colombo, Diego and Maathuis, Marloes H. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[8] Colombo, Diego, Maathuis, Marloes H, Kalisch, Markus, and Richardson, Thomas S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[9] Cui, Ruifei, Groot, Perry, and Heskes, Tom. Copula PC algorithm for causal discovery from mixed data. In *ECML PKDD*, pp. 377–392. Springer, 2016.

[10] Finney, Sara J and DiStefano, Christine. Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course*, pp. 269–314, 2006.

[11] Harris, Naftali and Drton, Mathias. PC algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research*, 14(Jan):3365–3383, 2013.

[12] Heutink, Peter, Verhuls, Frank C, and Boomsma, Dorret I. A longitudinal twin study on IQ, executive functioning, and attention problems during childhood and early adolescence. *Acta neurol. belg*, 106: 191–207, 2006.

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

[14] Kalisch, Markus and Bühlmann, Peter. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *The Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[15] Kummerfeld, Erich and Ramsey, Joseph. Causal clustering for 1-factor measurement models. In *SIGKDD*, pp. 1655–1664. ACM, 2016.

[16] Martínez-Torres, M Rocío. A procedure to design a structural and measurement model of intellectual capital: an exploratory study. *Information & Management*, 43(5):617–626, 2006.

[17] Murphy, Kevin P. Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1(2σ2):16, 2007.

[18] Murray, Jared S, Dunson, David B, Carin, Lawrence, and Lucas, Joseph E. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502): 656–665, 2013.

[19] Nøvik, Torunn Stene, Hervas, Amaia, Ralston, Stephen J, Dalsgaard, Søren, Pereira, Rob Rodrigues, Lorenzo, Maria J, Group, ADORE Study, et al. Influence of gender on attention-deficit/hyperactivity disorder in Europe–ADORE. *European child & adolescent psychiatry*, 15(1): i15–i24, 2006.

[20] Pearl, Judea. *Causality*. Cambridge university press, 2009.

[21] Roverato, Alberto. Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika*, 87(1): 99–112, 2000.

[22] Roverato, Alberto. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3): 391–411, 2002.

[23] Silva, Ricardo, Scheines, Richard, Glymour, Clark, and Spirtes, Peter. Learning measurement models for unobserved variables. In *UAI*, pp. 543–550, 2002.

[24] Silva, Ricardo, Scheines, Richard, Glymour, Clark, and Spirtes, Peter. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

[25] Skeem, Jennifer L and Cauffman, Elizabeth. Views of the downward extension: Comparing the youth version of the psychopathy checklist with the youth psychopathic traits inventory. *Behavioral sciences & the law*, 21(6):737–770, 2003.

[26] Sokolova, Elena, Groot, Perry, Claassen, Tom, van Hulzen, Kimm J, Glennon, Jeffrey C, Franke, Barbara, Heskes, Tom, and Buitelaar, Jan. Statistical evidence suggests that inattention drives hyperactivity/impulsivity in attention deficit-hyperactivity disorder. *PLoS one*, 11(10):e0165120, 2016.

[27] Spirtes, Peter, Glymour, Clark N, and Scheines, Richard. *Causation, prediction, and search*. MIT press, 2000.

[28] Tsamardinos, Ioannis, Brown, Laura E, and Aliferis, Constantin F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[29] Ullebø, Anne Karin, Breivik, Kyrre, Gillberg, Christopher, Lundervold, Astri J, and Posserud, Maj-Britt. The factor structure of ADHD in a general population of primary school children. *Journal of Child Psychology and Psychiatry*, 53(9):927–936, 2012.

[30] van Steijn, Daphne J, Richards, Jennifer S, Oerlemans, Anoek M, de Ruiter, Saskia W, van Aken, Marcel AG, Franke, Barbara, Buitelaar, Jan, Rommelse, Nanda NJ, et al. The co-occurrence of autism spectrum disorder and attention-deficit/hyperactivity disorder symptoms in parents of children with ASD or ASD with ADHD. *Journal of Child Psychology and Psychiatry*, 53(9):954–963, 2012.

[31] Willcutt, Erik G. The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics*, 9(3):490–499, 2012.

[32] Zhang, Kun, Zhang, Jiji, Huang, Biwei, Schölkopf, Bernhard, and Glymour, Clark. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *UAI*, 2016.