# Stable specification search in structural equation model with latent variables

Ridho Rahmadi[a,b,*], Perry Groot[b], Tom Heskes[b]

[a]*Department of Informatics, Universitas Islam Indonesia.*
[b]*Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands.*

## Abstract

In our previous study, we introduced stable specification search for cross-sectional data (S3C). It is an exploratory causal method that combines stability selection concept and multi-objective optimization to search for stable and parsimonious causal structures across the entire range of model complexities. In this study, we extended S3C to S3C-Latent, to model causal relations between latent variables. We evaluated S3C-Latent on simulated data and compared the results to those of PC-MIMBuild, an extension of the PC algorithm, the state-of-the-art causal discovery method. The comparison showed that S3C-Latent achieved better performance. We also applied S3C-Latent to real-world data of children with attention deficit/hyperactivity disorder and data about measuring mental abilities among pupils. The results are consistent with those of previous studies.

*Keywords:* causal modeling, structural equation model with latent variables, specification search, stability selection, multi-objective optimization

## 1. Introduction

In many empirical sciences, it is of great interest to identify causal relationships among entities that are not measured directly. Such entities are called latent variables or factors. A latent variable is typically used to represent a rather general concept that influences multiple measured or observed variables obtained from, for example, a questionnaire designed by experts in the field [1]. An example from psychometrics is the fatigue catastrophizing scale (FCS) [2], that is used to assess a patient's tendency of catastrophizing, a cognitive process characterized by a lack of confidence and an expectation of negative outcomes [3], as a response of experiencing fatigue. It is a 10-item scale, and each item therein, e.g., "I find myself expecting the worst when I am fatigued", is rated on 5-point scales (1 means *never true* and 5 means *all of the time true*).

Domains involving both observed and latent variables can be modeled using a structural equation model (SEM) with latent variables (also called a general SEM in [4]), that can be divided into two main components: the measurement model and the structural

---

[*]Corresponding author, email: r.rahmadi@cs.ru.nl, visiting address: Faculty of Science, Toernooiveld 212, 6525EC Nijmegen, the Netherlands.

model. The measurement model captures relationships between the latent variables and their corresponding observed variables, often called indicators. The structural model represents relationships among latent variables. A few approaches have been developed to discover the measurement model, such as factor analysis or a more advanced alternative introduced by [1]. Recently, a new method has been introduced to model causal relations between latent variables in the structural model [5].

To estimate underlying causal mechanisms is, however, a complex problem. First, given $p$ variables in the data, there will be $3^{p(p-1)/2}$ possible recursive structural models (described in Section 2.1)[6]. Thus, even a modest number of variables implies an immense number of possible models. A typical SEM procedure starts with a hypothesized model, scores the model, and realizes a few model refinements (often called a specification search) to improve the score [7], resulting in only a small number of model evaluations. Such a specification search is not intended to infer causal structures, but rather to detect and correct specification error between a proposed and the true model [8]. Second, model estimation is known to be notoriously unstable. That is, a slight fluctuation in the data can lead to a considerable change in the final model. This makes that typical approaches in literature, that are based solely on a single estimation, cannot robustly estimate a causal model.

In [9], we introduced stable specification search for cross-sectional data (S3C). S3C is an exploratory causal method that uses a (non-general) SEM to represent causal models with only observed variables. S3C represents a causal model using a SEM with observed variables. In the present study, we extend S3C to S3C-Latent to model causal relationships among latent variables. S3C and S3C-Latent are designed to resolve the aforementioned problems of an immense number of possible models and inherent instability in model estimation. We search for causal relations on the structural model and assume that the measurement model is given and pure, i.e., each indicator is related to only one latent variable [1]. In addition, we also allow demographic variables such as gender and age to be included in the structural model.

We evaluate the performance of S3C-Latent on different schemes of simulation and compare the results with those of PC-MIMBuild, an extension of the PC algorithm [10]. The results demonstrate that S3C-Latent performs better. We also apply S3C-Latent to real-world data from different domains, and the results are consistent with those of earlier studies. This paper is organized as follows. Section 2 describes SEMs with latent variables and the S3C procedure. Section 3 introduces the proposed method, S3C-Latent. Section 4 reports and discusses the results on simulated and real-world data. Finally, Section 5 gives the conclusions of this study.

## 2. Background

### 2.1. Structural equation model with latent variables

#### 2.1.1. Representation

We use a representation of a SEM with latent variables as described in [11, 4]. The SEM consists of the structural model that represents causal relationships among latent variables, and the measurement model that represents relationships from latent to observed variables. In the literature, the latent variable is often called a factor, and the

2

observed variable is often called an indicator, a manifest, or a proxy. In this paper, we use those terms interchangeably. The structural model reads

$$\boldsymbol{\eta} = \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta}, \tag{1}$$

where $\boldsymbol{\eta}$ is an $m \times 1$ vector of latent endogenous (effect) variables, $\boldsymbol{\xi}$ is an $n \times 1$ vector of latent exogenous (cause) variables, $\boldsymbol{\zeta}$ is an $m \times 1$ vector of disturbances on $\boldsymbol{\eta}$, $\boldsymbol{B}$ is an $m \times m$ matrix of coefficients among $\boldsymbol{\eta}$, and $\boldsymbol{\Gamma}$ is an $m \times n$ matrix of coefficients among $\boldsymbol{\xi}$. In addition, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ denote the covariance matrices of $\boldsymbol{\xi}$ and of $\boldsymbol{\zeta}$, respectively. We assume that $\mathbb{E}(\boldsymbol{\eta}) = \mathbb{E}(\boldsymbol{\xi}) = \mathbb{E}(\boldsymbol{\zeta}) = 0$, $\boldsymbol{\xi}$ uncorrelated with $\boldsymbol{\zeta}$, and that $(\boldsymbol{I} - \boldsymbol{B})$ is nonsingular. In this study, we assume a recursive structural model without reciprocal causal relations. Such a model can be represented using a directed acyclic graph (DAG) [12]. A DAG is a graph with a set of nodes $V$ and a set of directed edges $E$, in which no directed cycles (reciprocal) relations.

The measurement model represents influences from $\boldsymbol{\eta}$ to its observed variables, an $r \times 1$ vector $\boldsymbol{x}$, and from $\boldsymbol{\zeta}$ to its observed variables, a $q \times 1$ vector $\boldsymbol{y}$. The measurement model reads

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \\ \boldsymbol{y} &= \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}, \end{aligned} \tag{2}$$

where the $r \times n$ matrix $\boldsymbol{\Lambda}_x$ and $q \times m$ matrix $\boldsymbol{\Lambda}_y$ contain the structure coefficients associating latent variables and indicators, and the $r \times 1$ vector $\boldsymbol{\delta}$ and $q \times 1$ vector $\boldsymbol{\epsilon}$ contain errors on the indicators. In addition, an $r \times r$ matrix $\boldsymbol{\Theta}_\delta$ and a $q \times q$ matrix $\boldsymbol{\Theta}_\epsilon$ are the covariance matrices of $\boldsymbol{\delta}$ and of $\boldsymbol{\epsilon}$, respectively.

In general, a SEM procedure estimates a model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and evaluates how closely it matches the sample covariance matrix $\mathbf{S}$. The $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a function of model parameters $\boldsymbol{\theta}$ through

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{yx}(\boldsymbol{\theta}) \\ \boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) & \boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_y (\boldsymbol{I} - \boldsymbol{B})^{-1} (\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi}) \left[ (\boldsymbol{I} - \boldsymbol{B})^{-1} \right]' \boldsymbol{\Lambda}_y' + \boldsymbol{\Theta}_\epsilon, \tag{3}$$

$$\boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_x \boldsymbol{\Phi\Gamma}' \left[ (\boldsymbol{I} - \boldsymbol{B})^{-1} \right]' \boldsymbol{\Lambda}_y',$$

$$\boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_x \boldsymbol{\Phi\Lambda}_x' + \boldsymbol{\Theta}_\delta,$$

where $\boldsymbol{\Sigma}_{yy}(\boldsymbol{\theta})$ is a covariance matrix of the indicators $y$ written as a function of the parameters $\boldsymbol{\theta}$. There are analogous definitions for $\boldsymbol{\Sigma}_{yx}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}_{xy}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{xx}(\boldsymbol{\theta})$. The prime symbol indicates a matrix transpose. Figure 1 shows a SEM with three latent variables, each with three indicators.

*Measurement model with ordinal indicators*

The SEM with latent variables represented by Equations 1 and 2 assumes normally distributed indicators $\boldsymbol{x}$ and $\boldsymbol{y}$. Based on [13, 14], this model can be extended to ordinal indicators $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ by discretizing, e.g., $\boldsymbol{x}$ into $w$ categories through

$$\hat{x}_i = \hat{x}_{ij} \qquad \text{if } \tau_{j-1} \le x_i < \tau_j, \ j = 1, \dots, w, \ i = 1, \dots, r, \tag{4}$$
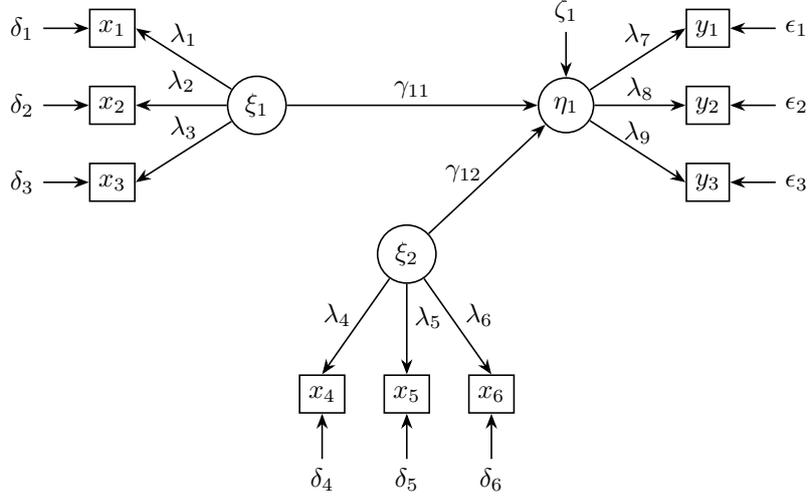
Figure 1: An example of a SEM with three latent variables, each with three indicators.

where $\tau_j$ is *threshold*, and for convenience, we define $\tau_0 = -\infty$ and $\tau_w = +\infty$. The threshold $\tau_j$ and the values of $\hat{x}_{ij}$ are assumed to be strictly increasing, i.e., $\tau_1 < \tau_2 < \ldots < \tau_{w-1}$, and $\hat{x}_{i1} < \hat{x}_{i2} < \ldots < \hat{x}_{iw}$. There are analogous expressions for discretizing $\boldsymbol{y}$ into $w$ categories.

We assume that $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are proxies of some continuous variables, and therefore use polychoric (between ordinal indicators) and polyserial (in the case of a mixture of continuous and ordinal indicators) correlations [13, 14], to form the matrix $\boldsymbol{S}$.

*Markov equivalence class*

A structural model represented by a DAG has its corresponding Markov equivalence class or often called a completed partially DAG (CPDAG) [15]. This implies that every probability distribution entailed by a structural model in a particular CPDAG, can also be derived by other structural models belonging to the same CPDAG.

*2.1.2. Identification and estimation*

An integral part of our procedure will be the estimation of (the parameters of) SEMs. These parameters can only be estimated when the so-called identification conditions are satisfied. For the SEMs with latent variables that we consider in this paper, these conditions are as follows [4, 16].

1. There are at least three or more indicators per latent variable.
2. Each row of $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ has only one nonzero element, i.e., an indicator cannot load on multiple latent variables (a pure model) [1].
3. Each latent variable is scaled, e.g., by setting one factor loading $\lambda_{ij}$ of each latent $\zeta_j$ to 1.
4. $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ is diagonal.

Condition 1 can be relaxed to latent variables with less than three indicators as follows. If there is a latent variable with two indicators, then the latent variable must have a

4

causal relation with other latent variables [4]. If there is a latent variable with only one indicator, then the corresponding indicator error is set to zero [16].

*Estimation*

After a SEM satisfies the identification conditions, its model parameters $\boldsymbol{\theta}$ can be estimated through a maximum likelihood procedure, by minimizing a fitting function that reads

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} F_{ML}(\boldsymbol{\theta}), \tag{5}$$

$$F_{ML}(\boldsymbol{\theta}) = \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \operatorname{Tr}\{\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\} - \log |\boldsymbol{S}| - p. \tag{6}$$

where $p$ is the number of observed variables, and $\boldsymbol{S}$ is the $p \times p$ sample covariance matrix of the observed variables.

*2.2. S3C procedure*

The S3C procedure can be divided into two main parts. The first part searches for the Pareto optimal causal models for the whole range of model complexities. This is realized by conducting a multi-objective optimization that mimics the idea of evolution [19]. The procedure is started with randomly generated SEMs that are scored based on two conflicting criteria: model fit and model complexity. The good models are characterized by those that fit the data well and have simple structures. They are selected to be parents to make new SEMs (children) in the next population of models by swapping and flipping model structures. This procedure is repeated many times, constituting many model refinements that lead to a better model population. At the end, the Pareto optimal models are obtained by filtering out those that are dominated by any other model in the population.

The second part of S3C aims to obtain SEM structures that are stable and simple (parsimonious). Adopting the idea of stability selection [20], S3C subsamples the data many times without replacement. For each of the subsets, S3C conducts the search that is described in the first part above, resulting in many Pareto optimal models. From those optimal models, S3C measures the edge and the causal path stability, that indicate the frequency of relations and the corresponding causal directions among pairs of variables across different model complexities (a precise definition of stability graphs is given in Appendix A). Finally, with some thresholds S3C infers stable and parsimonious model structures (called *relevant* structures) based on the edge and the causal path stability. For more details about the S3C procedure, we refer the readers to the original paper [9].

## 3. Proposed method

Originally, S3C is designed to model causal relationships among observed variables. In the present study we extend S3C to S3C-Latent, to model causal relationships among latent variables. In particular, we use the representation of SEMs with latent variables as described in Section 2.1. As the basic idea of S3C is to search through many possible models, there is no restriction that forbids a cause $\xi_i$ in one model to be an effect $\eta_i$ in another model, and vice versa. The same condition implies to the corresponding indicators, that is, $x_i$ in one model can be $y_i$ in another model, and vice versa. However, there could be an exemption from this rule if one intends to incorporate *prior knowledge*.

```
 1  Ensure identification conditions I fulfilled:
 2  if Λ indicates that any latent variable L_i ∈ L has indicators < 3 then
 3      if the number of indicators = 2 then
 4          Set a relation between L_i and one random latent L_j ∈ L
 5          Set one of the factor loadings on L_i to 1
 6      else
 7          Set the factor loading on L_i to 1
 8          Set the error on the indicator to 0
 9      end
10  else
11      Set one of the factor loadings on each L_i ∈ L to 1
12  end
13  Run S3C on D with information of L and satisfying C and I
```

Figure 2: Algorithm of S3C-Latent

In practice, prior knowledge on the domain of interest may exist, e.g., results of previous studies that leads to a constraint on a particular causal relation. For example, based on an earlier study on patients with medically unexplained fatigue, it is known that the objective physical activity does not reduce the level of fatigue directly. In terms of a SEM with latent variables, this prior knowledge can be translated into a structural model with no directed edge from latent variable $O$ (denoting the objective physical activity) to latent variable $F$ (denoting the level of fatigue). Note that a directed path from $O$ to $F$ is still allowed, e.g., a path $O \rightarrow \cdots \rightarrow F$ with any latent variables in between. S3C and S3C-Latent allow for incorporation of such prior knowledge.

In this study we further assume that the measurement model is given and pure. With respect to the data, we assume that samples have been generated independently and identically distributed (iid) from a linear Gaussian SEM.

Like S3C, to score a SEM, S3C-Latent uses the likelihood ratio test statistic [21] (equal to $(N-1)F_{ML}$ from Equation (6), with $N$ the sample size, which is often denoted $\chi_2$) to indicate the model fit, and the number of relations in the structural model (equals to the number of nonzero elements in matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ in Equation (1)), to indicate the model complexity. These scoring metrics are, however, not inherent to S3C-Latent, and can be replaced by different scoring metrics.

### 3.1. S3C-Latent

Let $D$ be the data set, $L = \{L_1, \ldots, L_n\}$ be a set of $n$ latent variables, $\Lambda$ be a matrix of factor loadings, and $C$ be prior knowledge. Figure 2 gives pseudocode of the S3C-Latent procedure.

Lines 1 to 11 are to ensure that the measurement model identification conditions described in Section 2.1.2 are fulfilled. In particular, Line 2 checks if there are any latent variables $L_i \in L$ having less than 3 indicators. If so, then for each $L_i$, Line 3 is realized or Line 11 otherwise. Line 3 checks whether the number of indicators of $L_i$ is 2 or 1. In the case of 2 indicators, S3C-Latent sets a relation between the latent variable $L_i$ and a random latent variable $L_j \in L$ ($L_i$ can be either a cause or an effect), and fixes

one of its factor loadings to 1 (Lines 4 and 5). In the case of 1 indicator, S3C-Latent sets the factor loading of $L_i$ to 1 and the indicator error to 0 (Lines 7 and 8). Line 11 is realized when all latent variables have at least 3 indicators. In this case, one of the factor loadings on each latent variable is set to 1. Finally, Line 13 runs S3C on data set $D$ with information of latent variables from $L$, satisfying any constraints in $C$, and fulfilling model identification conditions in $I$. By satisfying constraints in $C$ (if any), S3C-Latent ensures that all SEMs that are generated and refined, are consistent with the prior knowledge stated in $C$.

We implemented S3C-Latent as an R package and made it available online.[1] All of the source codes are provided, allowing interested users to modify the functionality.

## 4. Result and discussion

### 4.1. Application to simulated data

To simulate different practical cases, we conducted different schemes of simulation by generating data from models with different number of latent variables, sample size, number of indicators, and types of indicators, e.g., continuous and ordinal. All of the R scripts used and the data sets generated for this simulation are available at the same repository of the R package.

### 4.1.1. Model and data generation

First, we randomly generated structural models of $n$ latent variables, by generating DAGs of $n$ nodes with a probability $s = 2/(n-1)$ when relating a node (latent variable) to another node. Each DAG can be represented as an $n \times n$ adjacency matrix $\boldsymbol{A}$. As the matrix $\boldsymbol{A}$ captures causal relations among latent variables $L = \{L_1, \ldots, L_n\}$, the matrices $\boldsymbol{B}$ and $\boldsymbol{\Gamma}$ in Equation (1) are submatrices of $\boldsymbol{A}$. For each structural model, we then randomly formed two pure measurement models, where one has 3 to 5 indicators and the other one has 1 to 4 indicators per latent variable. This is realized by generating an $h \times n$ random matrix of factor loadings $\boldsymbol{\Lambda}$ for each structural model, where $h$ is the total number of indicators from $n$ latent variables. The matrices $\boldsymbol{\Lambda}_x$ and $\boldsymbol{\Lambda}_y$ in Equation (2) are submatrices of $\boldsymbol{\Lambda}$. With this procedure, we generated 20 SEMs with 4 and with 6 latent variables.

Second, we simulated data sets of sizes 400, 1000, and 2000 from each SEM with normal distribution, and then created the corresponding ordinal data sets by discretizing the values of the indicators into 2 to 7 ordered categories randomly. Taken together, we simulated 480 data sets on which the performance of S3C-Latent has been evaluated. Table 1 summarizes the schemes of the simulation. Scheme $C_{3-5}$, for example, applied S3C-Latent to the data sets of continuous variables generated from SEMs with 3 to 5 indicators per latent variable. Moreover, we also applied PC-MIMBuild [1] on the same data sets, and then compared the results. PC-MIMBuild originates from the PC algorithm [10], a state-of-the-art causal discovery method.

---

[1]https://github.com/rahmarid/S3C-Latent

Table 1: Types of indicators and the number of indicators per latent variable in different schemes of simulation.

| Scheme | Type of indicators | Number of indicators |
|--------|--------------------|----------------------|
| $C_{3-5}$ | Continuous | 3 to 5 |
| $C_{1-4}$ | Continuous | 1 to 4 |
| $O_{3-5}$ | Ordinal | 3 to 5 |
| $O_{1-4}$ | Ordinal | 1 to 4 |

### 4.1.2. Parameter settings

A parameter setting of S3C-Latent consists of the number of subsets to draw (S), the number of iterations (I), the number of models to evaluate (P), crossover probability (C), and mutation probability (M). For applications to data sets generated from SEMs with 4 latent variables, we set S = 25, I = 30, P = 50, C = 0.45, and M = 0.01. For applications to data sets generated from SEMs with 6 latent variables, we used the same parameter setting, except that we set I = 50 and P = 100.

For a fair comparison, we also drew 25 subsets in the applications of PC-MIMBuild (similar to that in [22]) and measured the edge and the causal path stability as in S3C-Latent. We set the significance level for PC-MIMBuild in testing conditional independence to 0.01.

### 4.1.3. Evaluation performance

We used the receiver operating characteristic (ROC) curve [23] to evaluate the performance of S3C-Latent and of PC-MIMBuild. Here, the true positive rate (TPR) and the false positive rate (FPR) are computed based on the CPDAG of the true model [15], while increasing the thresholds of stability. We measured the ROC for both the edge and the causal path stability. For example, a true positive means that a causal path with any length predicted by S3C-Latent or PC-MIMBuild actually exists in the CPDAG of the true model, while a false positive means that the predicted causal path does not exist in the CPDAG of the true model.

Each simulation scheme comprises 20 data sets of sizes 400, 1000, and 2000, making it in total 60 data sets. We measured the area under the curve (AUC) of each ROC and then computed the mean of the AUCs that S3C-Latent obtained over 20 data sets (of the same sample size) and compared to that of PC-MIMBuild.

### 4.1.4. Discussion

Figures 3 and 4 show the plots of the mean AUC and the corresponding standard errors obtained by S3C-latent (black-solid lines) and PC-MIMBuild (blue-dashed lines), as a function of sample size. The top panels show results from Schemes $C_{3-5}$ and $O_{3-5}$, and the bottom panels show those from Schemes $C_{1-4}$ and $O_{1-4}$.

In all simulation schemes with 4 latent variables, S3C-Latent achieved higher AUCs than those achieved by PC-MIMBuild. In a similar trend, S3C-Latent obtained higher AUCs than those achieved by PC-MIMBuild on most simulation schemes with 6 latent variables. The discrepancy of the results of S3C-Latent and of PC-MIMBuild becomes clearer in the case of continuous indicators. Based on the results on simulations with 4 and 6 latent variables, the performance of S3C-Latent seems to be more robust compared to that of PC-MIMBuild across different sample sizes. In particular for small sample
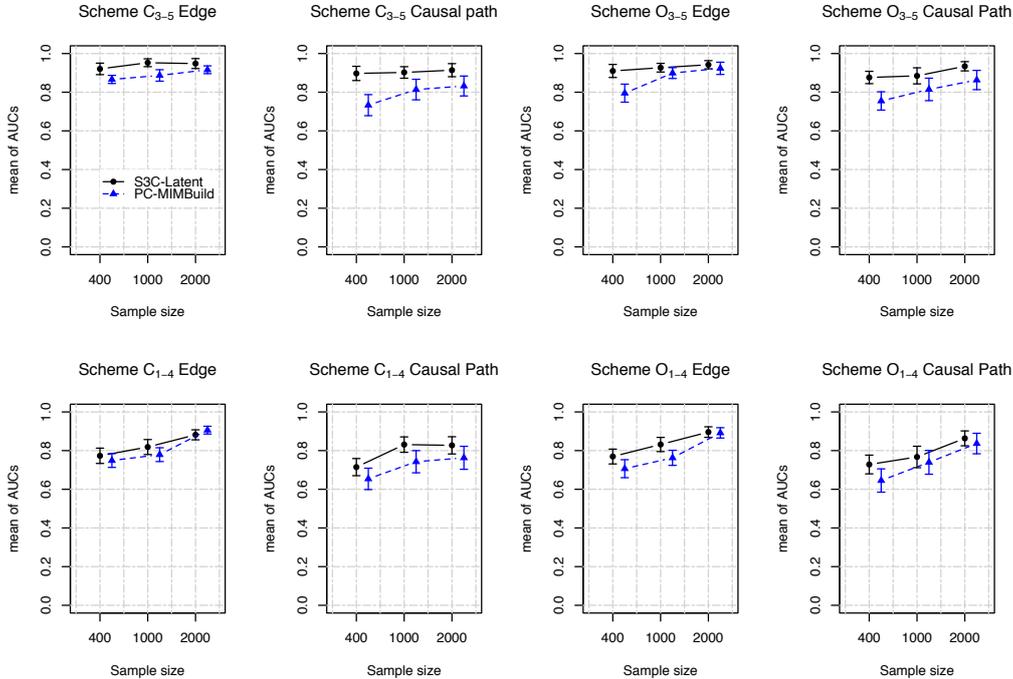
Figure 3: The mean AUC obtained by S3C-Latent (black-solid lines) and PC-MIMBuild (blue-dashed lines) over 20 data sets generated from SEMs with 4 latent variables, as a function of sample size. "Edge" indicates AUCs measured from the edge stability and "Causal Path" indicates those measured from the causal path stability. The plots in the top panels are results of simulations on the data sets generated from SEMs with 3 to 5 continuous (Scheme $C_{3-5}$) or ordinal (Scheme $O_{3-5}$) indicators per latent variable. The plots in the bottom panels are results of simulations on the data sets generated from SEMs with 1 to 4 continuous (Scheme $C_{1-4}$) or ordinal (Scheme $O_{1-4}$) indicators per latent variable.

sizes, typical to many real-world applications, S3C Latent appears to be more robust than PC-MIMBuild.

Figure B.9 in Appendix B shows comparisons of S3C-Latent's results in different schemes. As expected, in general the performance of S3C-Latent on data generated from SEMs with 3 to 5 indicators per latent variable is better than that of on data generated from SEMs with 1 to 4 indicators per latent variable.

### 4.2. Application to real-world data

#### 4.2.1. Attention deficit/hyperactivity disorder data

For the first application, we applied S3C-Latent to a data set about attention deficit/hyperactivity disorder (ADHD) that was collected in [24]. The data set comprises observations on the children, among which 236 have ADHD and 406 belong to the control group. From all of the samples, 269 are girls. There are four observed variables: gender, age, verbal intelligence quotient (VIQ), and performance intelligence quotient (PIQ). Moreover, there are 18 questions distributed into three groups to indicate three latent variables, namely inattention (questions 1-9), hyperactivity (questions 10-14), and
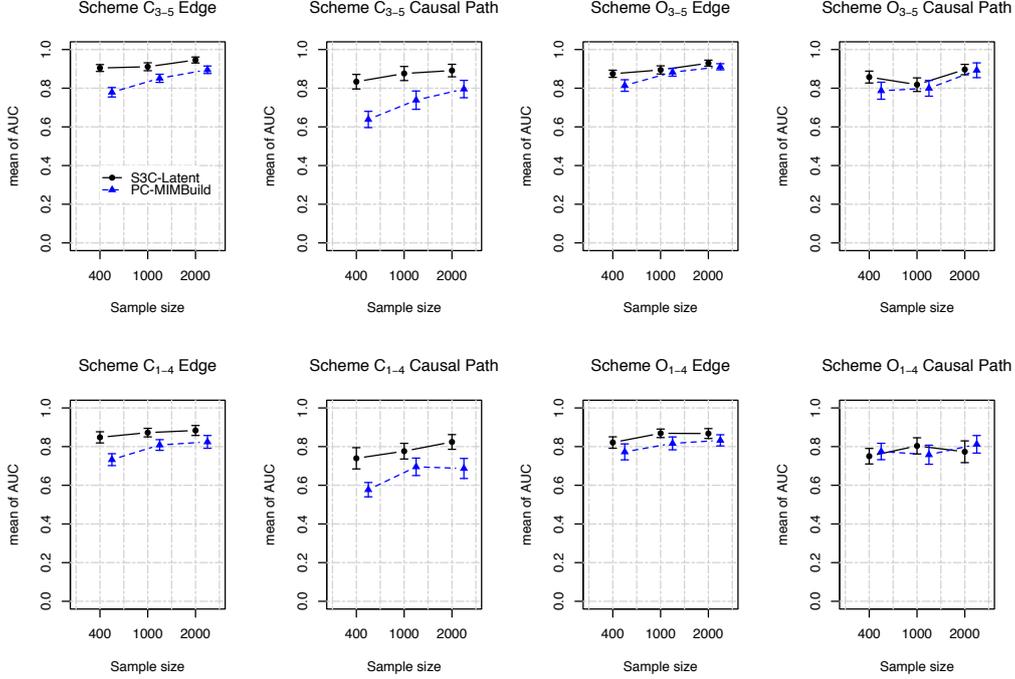
9

Figure 4: The mean AUC obtained by S3C-Latent (black-solid lines) and PC-MIMBuild (blue-dashed lines) over 20 data sets generated from SEMs with 6 latent variables, as a function of sample size. "Edge" indicates AUCs measured from the edge stability and "Causal Path" indicates those measured from the causal path stability. The plots in the top panels are results of simulations on the data sets generated from SEMs with 3 to 5 continuous (Scheme $C_{3-5}$) or ordinal (Scheme $O_{3-5}$) indicators per latent variable. The plots in the bottom panels are results of simulations on the data sets generated from SEMs with 1 to 4 continuous (Scheme $C_{1-4}$) or ordinal (Scheme $O_{1-4}$) indicators per latent variable.

impulsivity (questions 15-18). We intend to model the causal relations between those four observed and the three latent variables. The missing data, that are assumed to be missing at random (MAR), are about 0.78% and were imputed using expectation maximization (EM) [25]. We combined the imputation with our subsampling, resulting in different imputations across data subsets. The parameter settings of S3C-Latent are S = 100, I = 100, P = 200, C = 0.45, and M = 0.01. We included a prior knowledge that nothing causes gender.

### 4.2.2. Discussion on ADHD result

Figure 5 depicts the edge and the causal path stability graphs, with the $x$-axis indicating the model complexity and the $y$-axis indicating the selection probability. The threshold $\pi_{sel}$ (horizontal line) is set to 0.6 and indicates that any model structure with selection probability equal to or greater than $\pi_{sel}$ is considered stable [20]. The threshold $\pi_{bic}$ (vertical line) is set to the model complexity at which the minimum median of the BIC scores is found [9], which in this case was at 12, indicating that any model structure with model complexity equal to or lesser than $\pi_{bic}$ is considered simple or par-

simonious. Thus, the relevant model structures (stable and parsimonious) are those that pass through the top-left region of the stability graphs. The relevant model structures are then visualized with the steps described in [9], resulting in the graph shown in Figure 6. A more detailed description on how to interpret the graph is given in the figure caption.
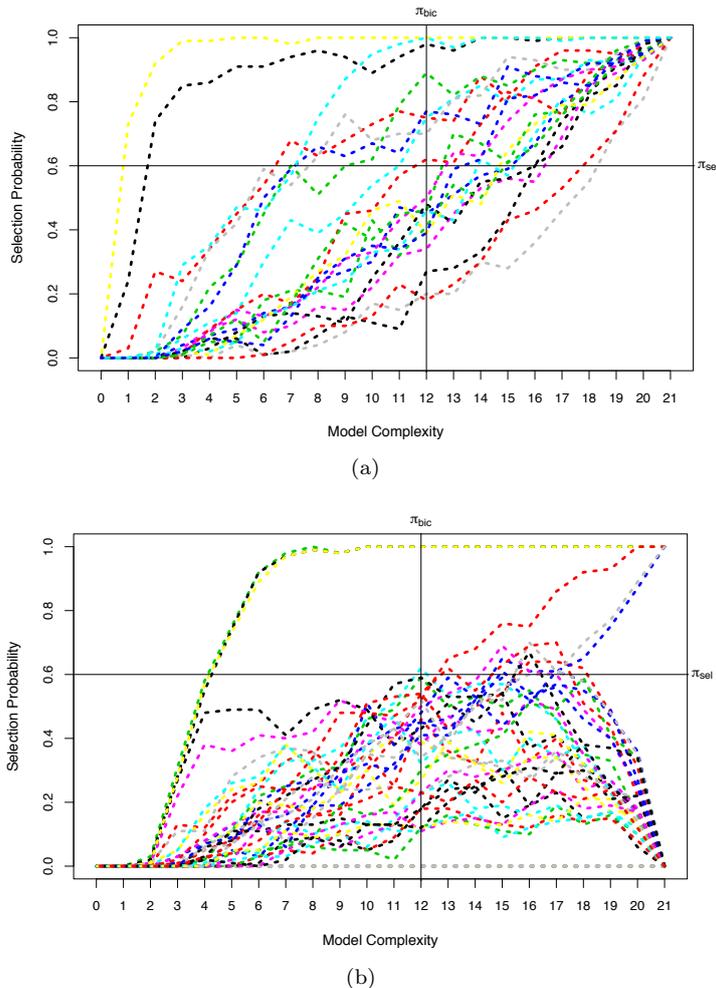


(a)



(b)

Figure 5: The (a) edge and the (b) causal path stability graphs from the ADHD application. Each line indicates the frequency of relations between a pair of variables across different model complexities. The frequency is computed relative to the number of SEMs in each model complexity. The causal path stability only takes into account causal relations with any length, while the edge stability takes into account all relations regardless of the direction.

Based on Figure 6, gender is found to influence inattention, hyperactivity, and impulsivity. This is in accordance with the studies of [26, 27, 28]which exhibited that boys with ADHD tend to have higher symptom levels. Meta-analyses in population-based [29] and clinically referred samples [30] suggested that boys are more likely to meet the
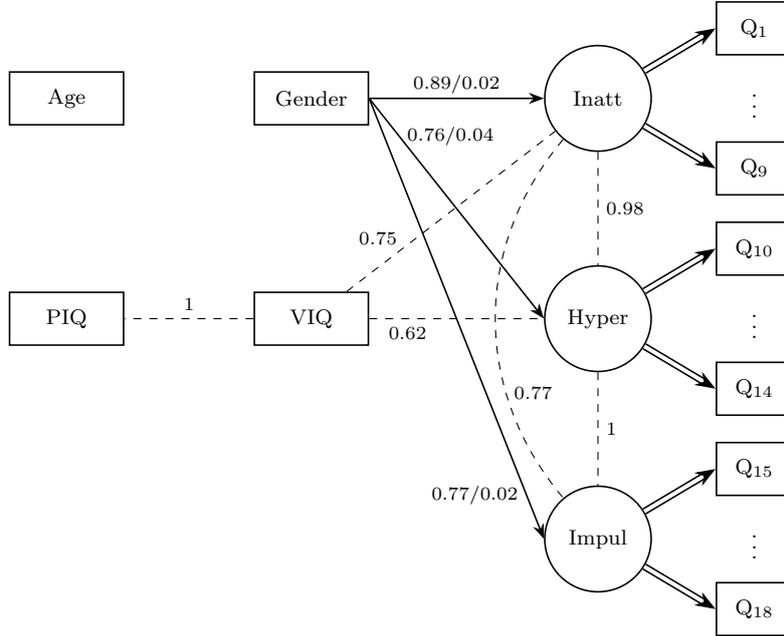
11

Figure 6: A dashed undirected edge exhibits strong association where the causal direction cannot be determined from the data. A directed edge (arrow) with a single line indicates a causal relation and that with a double line indicates a factor loading. All edges, except the factor loadings, are annotated with a reliability score, i.e., the highest selection probability an edge has across the relevant region of the edge stability graph. In addition, all arrows are annotated with an estimate of the total causal effect (see Appendix C for more detail). For example, 0.89/0.02 indicates a reliability score of 0.89 and a total causal effect of 0.02.

Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria for ADHD than girls.

VIQ is found associated with inattention and with hyperactivity, and that matches the studies of [31, 32] which reported that children with ADHD have significantly lower VIQ compared to children without ADHD. The relation between VIQ and PIQ is expected as both constitute an assessment of the subject's intelligence.

The association between inattention and hyperactivity is likely due to the direct causal relations from gender to both variables. A similar relation was found in the studies of [33, 34] which exhibited that inattention influences hyperactivity. S3C-Latent found that the causal path stability from inattention to hyperactivity is about 0.4. Having more samples might increase the stability (e.g., $> 0.6$), so as to justify the same causal direction.

The relation between hyperactivity and impulsivity (as also found by [33]) is also likely due to the direct causal relations from gender to both variables. In fact, in most studies hyperactivity and impulsivity are regarded as one combined feature.

The association between inattention and impulsivity possibly follows from the direct relations between inattention and impulsivity with hyperactivity. This association was also indicated by [33]. The associations between inattention, hyperactivity, and impul-

sivity are sensible as the three constitute the ADHD symptom. In addition, S3L did not find any relation between age with other variables.

### 4.2.3. Holzinger data

For the second application, we considered a well-known data set in the psychometric domain [35], obtained from an R package called MBESS 4.4.3 [36]. The aim of the study was to learn factor pattern for mental abilities among two groups of pupils with different biological and cultural backgrounds [37]. For this purpose, 301 students (155 females) from grades 7 and 8 were given 24 tests, measuring spatial, verbal, mental speed, memory, and mathematical ability. More details about the tests can be obtained from the package documentation. Our interest is to model causal relations among those latent variables including gender. We include a prior knowledge that nothing causes gender. We set the S3C-Latent parameters S = 100, I = 80, P = 150, C = 0.45, and M = 0.01.

### 4.2.4. Discussion on Holzinger result

Figure 7 displays the edge and the causal path stability graphs. We set $\pi_{sel} = 0.6$ and found that $\pi_{bic} = 11$ (cf. Section 4.2.2). Figure 8 visualizes the relevant structures into a graph.

Based on Figure 8, we see that gender influences spatial, mental speed, and memory ability. The causal relation from gender to spatial ability matches the meta-analysis of [38] that indicated that males tend to perform better than females in specific types of spatial ability. The causal relation from gender to memory is in agreement with the results of [39] and [40] that exhibited gender differences in favor of males on object and word location memory, and working memory index, respectively. Gender was also found to influence the mental speed, as indicated by [41] that pointed out a male advantage in reaction times using various measures.

It was also found that mathematical ability affects spatial ability. Previous studies indicated the association between the two [42, 43, 44]. Other studies emphasized that mathematics is one of the concepts that is mentally represented in a spatial format [45, 46]. Moreover, the study by [47] detailed specific mathematical tasks that predicts the most variance in the spatial ability of children aged 5 to 13.

All latent variables seem to be associated directly or indirectly. This might stem from some items on different latent variables that are possibly overlapping, for example, arithmetic (mathematical ability) and addition tasks (mental speed). Some associations are likely due to a common direct relation that a pair of variables has, for example, spatial ability and mental speed are affected by gender and both are associated. That being said, previous studies [48, 49, 50, 51, 52, 53] indicated and discussed most of the associations.

## 5. Conclusion and future work

It is often of great interest in many fields such as sociology, psychology, and medicine to model causal relationship among latent constructs that are indicated through observed proxies. In the present study, we extended S3C to S3C-Latent to model causal relations among latent variables. The chief aim of S3C-Latent is to resolve the problems of the inherent instability in model estimation and the immense number of possible models. To
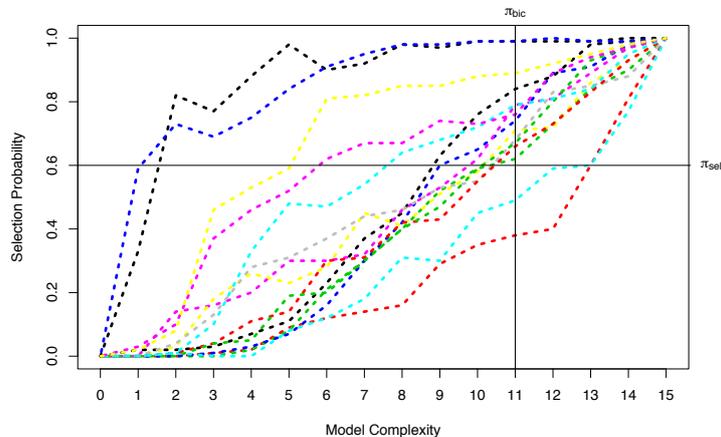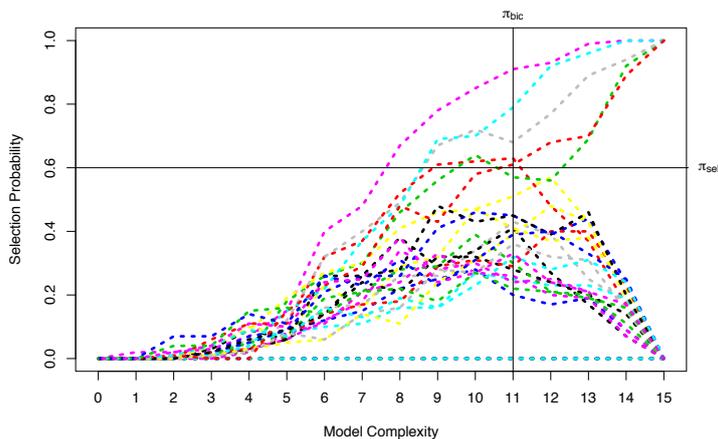
Figure 7: The (a) edge and the (b) causal path stability graphs from Holzinger application. Each line indicates the frequency of relations between a pair of variables across different model complexities. The frequency is computed relative to the number of SEMs in each model complexity. The causal path only stability takes into account causal relations with any length, while the edge stability takes into account all relations regardless of the direction.

realize that, S3C-Latent recasts the concept of stability selection into a multi-objective optimization problem, and jointly optimizes across the whole range of model complexities, resulting in Pareto optimal models. One also can see S3C-Latent as a general framework that can be integrated with other causal discovery methods for latent variables without altering their original assumptions. In principle, it can be realized by running a causal discovery method on different data subsets and then measuring the same edge and causal path stability based on the inferred models.

We compared the results of S3C-Latent to those of PC-MIMBuild on simulated data
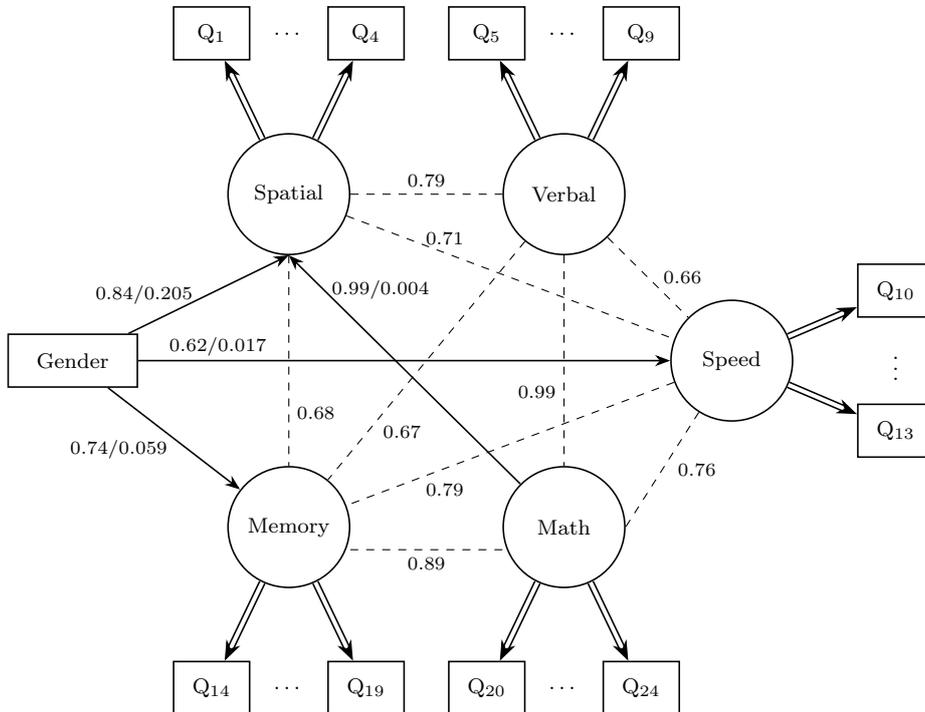
Figure 8: A dashed undirected edge exhibits strong association where the causal direction cannot be determined from the data. A directed edge (arrow) with a single line indicates a causal relation and that with a double line indicates a factor loading. All edges, except the factor loadings, are annotated with a reliability score, i.e., the highest selection probability an edge has across the relevant region of the edge stability graph. In addition, all arrows are annotated with an estimate of the total causal effect (see Appendix C for more detail). For example, 0.62/0.017 indicates a reliability score of 0.62 and a total causal effect of 0.017.

generated with different schemes, varying the number of latent variables, sample size, number of indicators, and types of indicators (continuous and ordinal). The comparison showed that S3C-Latent performs better than PC-MIMBuild. Moreover, we applied S3C-Latent to real-world ADHD and Holzinger data sets. In general, the results are consistent with those of earlier studies.

The current version of S3C-Latent, however, is designed for cross-sectional data, assumes no reciprocal causal relation, and cannot handle missing values other than through imputation prior to application. Thus, future work should consider extensions to longitudinal data, reciprocal causal relations, and a more sophisticated way of handling missing values.

## References

## References

[1] R. Silva, R. Scheine, C. Glymour, P. Spirtes, Learning the structure of linear latent variable models, Journal of Machine Learning Research 7 (Feb) (2006) 191–246.

[2] P. Jacobsen, L. Azzarello, D. Hann, Relation of catastrophizing to fatigue severity in women with breast cancer, Cancer Res Ther Control 8 (155) (1999) 1999–164.

[3] M. J. Sullivan, J. L. D'Eon, Relation between catastrophizing and depression in chronic pain patients., Journal of abnormal psychology 99 (3) (1990) 260.

[4] K. Bollen, Structural Equations With Latent Variables, WILEY, 1989.

[5] R. Cui, P. Groot, M. Schauer, T. Heskes, Learning the causal structure of copula models with latent variables, in: Conference on Uncertainty in Artificial Intelligence (Accepted), 2018.

[6] S. Harwood, R. Scheines, Genetic algorithm search over causal models.

[7] J. S. Long, Covariance structure models: An introduction to LISREL, no. 34, Sage, 1983.

[8] R. MacCallum, Specification searches in covariance structure modeling, Psychological Bulletin 100 (1) (1986) 107–120.

[9] R. Rahmadi, P. Groot, M. Heins, H. Knoop, T. Heskes, the OPTIMISTIC consortium, Causality on cross-sectional data: Stable specification search in constrained structural equation modeling, Applied Soft Computing 52 (2017) 687–698.

[10] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, Vol. 81, MIT press, 2000.

[11] K. G. Joreskog, Structural equation models in the social sciences: Specification estimation and testing, Applications of statistics (1977) 265–287.

[12] P. Spirtes, Introduction to causal inference, The Journal of Machine Learning Research 11 (2010) 1643–1662.

[13] U. Olsson, F. Drasgow, N. J. Dorans, The polyserial correlation coefficient, Psychometrika 47 (3) (1982) 337–347.

[14] F. Drasgow, Polychoric and polyserial correlations, Encyclopedia of statistical sciences.

[15] D. M. Chickering, Learning equivalence classes of Bayesian-network structures, The Journal of Machine Learning Research 2 (2002) 445–498.

[16] D. Kenny, D. Kashy, N. Bolger, Data analysis in social psychology (in d. gilbert, s. fiske, & g. lindzey (eds.). the handbook of social psychology (vol. 1, pp. 233–265) (1998).

[17] R. Weston, P. A. Gore Jr, A brief guide to structural equation modeling, The counseling psychologist 34 (5) (2006) 719–751.

[18] K. G. Jöreskog, Structural analysis of covariance and correlation matrices, Psychometrika 43 (4) (1978) 443–477.

[19] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation 6 (2) (2002) 182–197.

[20] N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72 (4) (2010) 417–473.

[21] K. G. Jöreskog, Some contributions to maximum likelihood factor analysis, Psychometrika 32 (4) (1967) 443–482.

[22] J. Ramsey, Bootstrapping the PC and CPC algorithms to improve search accuracy.

[23] T. Fawcett, ROC graphs: Notes and practical considerations for researchers, Machine learning 31 (2004) 1–38.

[24] D. J. van Steijn, J. S. Richards, A. M. Oerlemans, S. W. de Ruiter, M. A. van Aken, B. Franke, J. Buitelaar, N. N. Rommelse, 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) (2012) 954–963.

[25] J. Honaker, G. King, M. Blackwell, et al., Amelia II: A program for missing data, Journal of statistical software 45 (7) (2011) 1–47.

[26] M. Gaub, C. L. Carlson, Gender differences in ADHD: a meta-analysis and critical review, Journal of the American Academy of Child & Adolescent Psychiatry 36 (8) (1997) 1036–1045.

[27] J. Gershon, J. Gershon, A meta-analytic review of gender differences in ADHD, Journal of attention disorders 5 (3) (2002) 143–154.

[28] J. J. Bauermeister, P. E. Shrout, L. Chávez, M. Rubio-Stipec, R. Ramírez, L. Padilla, A. Anderson, P. García, G. Canino, ADHD and gender: are risks and sequela of ADHD the same for boys and girls?, Journal of Child Psychology and Psychiatry 48 (8) (2007) 831–839.

[29] E. G. Willcutt, B. F. Pennington, J. C. DeFries, Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties, Journal of Abnormal Child Psychology 28 (2) (2000) 149–159.

[30] T. S. Nøvik, A. Hervas, S. J. Ralston, S. Dalsgaard, R. R. Pereira, M. J. Lorenzo, A. S. Group, et al., Influence of gender on attention-deficit/hyperactivity disorder in europe–adore, European Child & Adolescent Psychiatry 15 (1) (2006) i15–i24.

[31] G. Andreou, P. Agapitou, A. Karapetsas, Verbal skills in children with ADHD, European Journal

of Special Needs Education 20 (2) (2005) 231–238.

[32] T. W. Frazier, H. A. Demaree, E. A. Youngstrom, Meta-analysis of intellectual and neuropsychological test performance in attention-deficit/hyperactivity disorder., Neuropsychology 18 (3) (2004) 543.

[33] E. Sokolova, A. M. Oerlemans, N. N. Rommelse, P. Groot, C. A. Hartman, J. C. Glennon, T. Claassen, T. Heskes, J. K. Buitelaar, A causal and mediation analysis of the comorbidity between attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD), Journal of autism and developmental disorders 47 (6) (2017) 1595–1604.

[34] E. Sokolova, P. Groot, T. Claassen, K. J. van Hulzen, J. C. Glennon, B. Franke, T. Heskes, J. Buitelaar, Statistical evidence suggests that inattention drives hyperactivity/impulsivity in attention deficit-hyperactivity disorder, PLoS one 11 (10) (2016) e0165120.

[35] K. J. Holzinger, F. Swineford, A study in factor analysis: The stability of a Bi-factor solution., Supplementary Educational Monographs.

[36] K. Kelley, Methods for the behavioral, educational, and social sciences: An R package, Behavior Research Methods 39 (4) (2007) 979–984.

[37] D. E. Scates, A study in factor analysis: The stability of a Bi-Factor solution. Karl J. Holzinger, Frances Swineford, The Elementary School Journal 40 (6) (1940) 468–469. doi:10.1086/457783.

[38] M. C. Linn, A. C. Petersen, Emergence and characterization of sex differences in spatial ability: A meta-analysis, Child development (1985) 1479–1498.

[39] Z. Cattaneo, A. Postma, T. Vecchi, Gender differences in memory for object and word locations, Quarterly Journal of Experimental Psychology 59 (5) (2006) 904–919.

[40] R. S. Longman, D. H. Saklofske, T. S. Fung, Wais-iii percentile scores by education and sex for us and canadian populations, Assessment 14 (4) (2007) 426–432.

[41] G. Der, I. J. Deary, Age and sex differences in reaction time in adulthood: results from the united kingdom health and lifestyle survey., Psychology and aging 21 (1) (2006) 62.

[42] S. A. Burnett, D. M. Lane, L. M. Dratt, Spatial visualization and sex differences in quantitative ability, Intelligence 3 (4) (1979) 345–354.

[43] M. B. Casey, R. L. Nuttall, E. Pezaris, Spatial-mechanical reasoning skills versus mathematics self-confidence as mediators of gender differences on mathematics subtests using cross-national gender-based items, Journal for research in mathematics education (2001) 28–57.

[44] N. M. Robinson, R. D. Abbott, V. W. Berninger, J. Busse, Structure of abilities in math-precocious young children: Gender similarities and differences., Journal of Educational Psychology 88 (2) (1996) 341.

[45] L. W. Barsalou, Grounded cognition, Annu. Rev. Psychol. 59 (2008) 617–645.

[46] G. Lakoff, R. E. Núñez, Where mathematics comes from: How the embodied mind brings mathematics into being, AMC 10 (2000) 12.

[47] K. S. Mix, S. C. Levine, Y.-L. Cheng, C. Young, D. Z. Hambrick, R. Ping, S. Konstantopoulos, Separate but correlated: The latent structure of space and mathematics across development., Journal of Experimental Psychology: General 145 (9) (2016) 1206.

[48] R. Lynn, The intelligence of the mongoloids: A psychometric, evolutionary and neurological theory, Personality and individual differences 8 (6) (1987) 813–844.

[49] J. Cantor, R. W. Engle, G. Hamilton, Short-term memory, working memory, and verbal abilities: How do they relate?, Intelligence 15 (2) (1991) 229–246.

[50] B. De Smedt, R. Janssen, K. Bouwens, L. Verschaffel, B. Boets, P. Ghesquière, Working memory and individual differences in mathematics achievement: A longitudinal study from first grade to second grade, Journal of experimental child psychology 103 (2) (2009) 186–201.

[51] M. Daneman, P. A. Carpenter, Individual differences in working memory and reading, Journal of verbal learning and verbal behavior 19 (4) (1980) 450–466.

[52] R. K. Vukovic, N. K. Lesaux, The relationship between linguistic skills and arithmetic knowledge, Learning and Individual Differences 23 (2013) 87–91.

[53] M. Mitolo, S. Gardini, P. Caffarra, L. Ronconi, A. Venneri, F. Pazzaglia, Relationship between spatial ability, visuospatial working memory and self-assessed spatial orientation ability: a study in older adults, Cognitive processing 16 (2) (2015) 165–176.

[54] M. H. Maathuis, M. Kalisch, P. Bühlmann, et al., Estimating high-dimensional intervention effects from observational data, The Annals of Statistics 37 (6A) (2009) 3133–3164.

[55] J. Pearl, Causality: models, reasoning and inference, Vol. 29, Cambridge University Press, 2000.

[56] Z. Ghahramani, G. E. Hinton, et al., The EM algorithm for mixtures of factor analyzers, Tech. rep., Technical Report CRG-TR-96-1, University of Toronto (1996).

## Appendix A. Stability graphs

**Definition 1.** (Stability graphs [9]) Let $A$ and $B$ be two variables and $G$ a multiset (or bag) of CPDAGS. Let $G_c$ be the submultiset of $G$ containing all CPDAGS with complexity $c$. The edge stability for $A$ and $B$ at complexity $c$ is the number of models in $G_c$ for which there exists an edge between $A$ and $B$ (i.e., $A \to B$, $B \to A$, or $A - B$) divided by the total number of models in $G_c$. The causal path stability for $A$ to $B$ at complexity $c$ is the number of models in $G_c$ for which there is a directed path from $A$ to $B$ (of any length) divided by the total number of models in $G_c$. The terms edge stability graph and causal path stability graph are used to denote the corresponding measures for all variable pairs and all complexity levels.

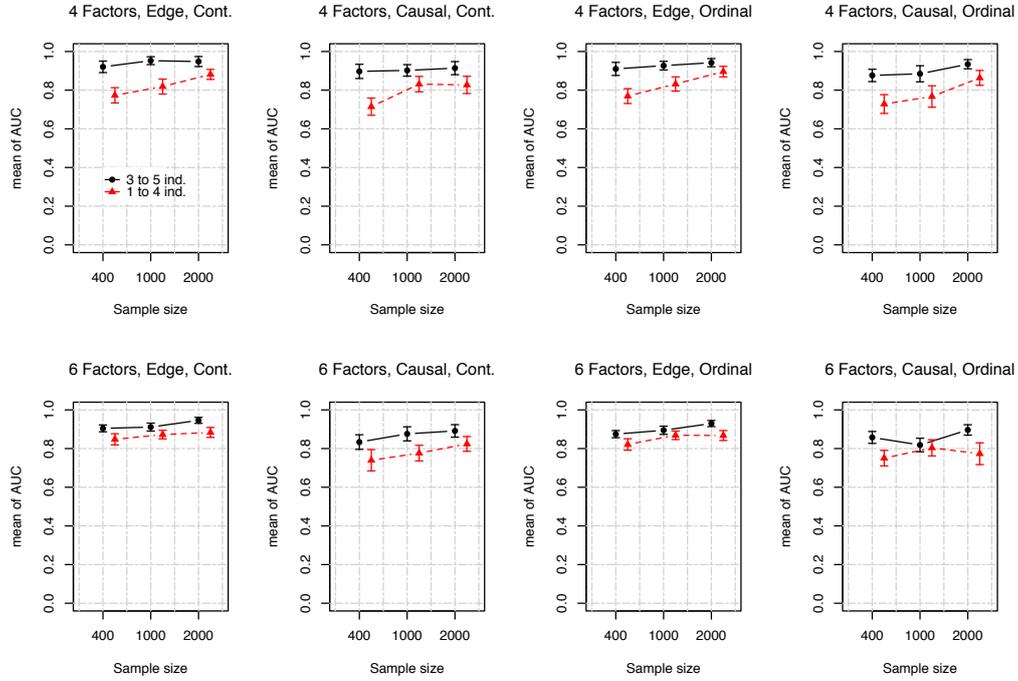## Appendix B. Comparison of S3C-Latent results



Figure B.9: Comparison between the mean AUC obtained by S3C-Latent on the data generated from SEMs with 3 to 5 indicators per latent variable to those obtained on data generated from SEMs with 1 to 4 indicators per latent variable. The plots in the top panel are results of simulation on the data generated from SEMs of 4 latent variables and those in the bottom panel are results of simulation on the data sets generated from SEMs of 6 latent variables.

## Appendix C. Causal effect estimation

The procedure originates from IDA method[54]. Let $G = \{G_1, \ldots, G_m\}$ be a CPDAG to which $m$ DAGs belong to. For each DAG $G_j$, IDA administers intervention calcu-

lus [55] to obtain multisets $\Theta_i = \{\theta_{ij}\}_{j\in 1,...,m}, i = 1, \ldots, p$, where $p$ is the number of covariates. In particular, $\theta_{ij}$ defines the possible causal effect of $X_i$ on $Y$ in graph $G_j$. Assuming that the data have been drawn from a linear Gaussian model, the causal effects of $X_i$ on $Y$ can be estimated through a regression of $Y$ on $X_i$ and its parents. Note that the possible causal effect $\theta_{ij}$ can be direct, indirect, or addition of the both effects, depending on the structure of $G_j$.

As S3C and S3C-Latent result in many CPDAGs representing optimal models, we extend IDA as follows. We particularly use $G_{\pi_{\mathrm{bic}}}$, the CPDAGs of all optimal models with complexity equal to $\pi_{\mathrm{bic}}$. For each CPDAG $G \in G_{\pi_{\mathrm{bic}}}$, we compute the possible causal effects of each relevant causal path, e.g., $X$ on $Y$, obtaining estimates $\Theta_{X\to Y}^k, k = 1, \ldots, N$, where $N$ is the number of data subsets. We then concatenate all causal effect estimates in $\Theta_{X\to Y}^k$ into a single multiset $\Theta_{X\to Y}$.

Finally, we use the median $\tilde{\Theta}_{X\to Y}$ to represent the estimated total causal effects from $X$ to $Y$. In the case of both $X$ and $Y$ are continuous, $\tilde{\Theta}_{X\to Y}$ is standardized via $\tilde{\Theta}_{X\to Y} \times \sigma_X/\sigma_Y$, where $\sigma_X$ and $\sigma_Y$ are the standard deviations of the covariate and the response, respectively.

In order to estimate causal effect from a SEM with latent variables, we need to sample data from the latent variables. The steps to sample from latent variables are described in Appendix D.

## Appendix D. Sampling from latent variable

In what follows, we describe the steps for the sampling. Suppose we are given a measurement model that reads

$$\boldsymbol{x} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{D.1}$$

where $\boldsymbol{\eta}$ is a vector of latent (for simplicity, can be either endogenous or exogenous) variables, $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{x}$ is a vector of indicators, $\boldsymbol{\epsilon}$ contain errors of the indicators, and $\boldsymbol{\Theta}$ is a covariance matrix of $\boldsymbol{\epsilon}$ and is diagonal.

Following [56], given $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$, the expected value of latent variables $\boldsymbol{\eta}$ can be computed via the linear projection:

$$\mathbb{E}(\boldsymbol{\eta}|\boldsymbol{x}) = \beta\boldsymbol{x}, \tag{D.2}$$

with $\beta \equiv \boldsymbol{\Lambda}'(\boldsymbol{\Theta} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1}$, and the variance of $\boldsymbol{\eta}$ can be computed through,

$$\mathrm{Var}(\boldsymbol{\eta}|\boldsymbol{x}) = I - \beta\boldsymbol{\Lambda}. \tag{D.3}$$

We can then sample $\hat{\boldsymbol{\eta}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, where $\boldsymbol{\mu}$ is the mean of vector $\mathbb{E}(\boldsymbol{\eta}|\boldsymbol{x})$ and $\boldsymbol{\sigma}^2 = \mathrm{Var}(\boldsymbol{\eta}|\boldsymbol{x})$.