This is a “preproof” accepted article for *Psychometrika*.
This version may be subject to change during the production process.
DOI: 10.1017/psy.2024.11

Casual Structural Modeling of Survey Questionnaires via a Bootstrapped Ordinal Bayesian Network Approach

Yang Ni[1,2,*] and Su Chen[2,3,4] and Zeya Wang[5]
[1]Department of Statistics, Texas A&M University
[2]Department of Statistics and Data Sciences, The University of Texas at Austin
[3]Center of Transforming Data to Knowledge, Rice University
[4]Department of Statistics, Rice University
[5]Dr. Bing Zhang Department of Statistics, University of Kentucky
[*]Correspondence: yni@stat.tamu.edu

Competing Interests

The authors declare no competing interests.

Data Availability

The dataset analyzed in this paper is available in the online supplementary materials of the following paper:

McNally, Richard J., Patrick Mair, Beth L. Mugno, and Bradley C. Riemann. "Co-morbid obsessive–compulsive disorder and depression: A Bayesian network approach." Psychological Medicine 47, no. 7 (2017): 1204-1214.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Abstract

Survey questionnaires are commonly used by psychologists and social scientists to measure various latent traits of study subjects. Various causal inference methods such as the potential outcome framework and structural equation models have been used to infer causal effects. However, the majority of these methods assume the knowledge of true causal structure, which is unknown for many applications in psychological and social sciences. This calls for alternative causal approaches for analyzing such questionnaire data. Bayesian networks are a promising option as they do not require causal structure to be known *a priori* but learn it objectively from data. Although we have seen some recent successes in using Bayesian networks to discover causality for psychological questionnaire data, their techniques tend to suffer from causal non-identifiability with observational data. In this paper, we propose using a state-of-the-art Bayesian network that is proven to be fully identifiable for observational ordinal data. We develop a causal structure learning algorithm based on an asymptotically justified BIC score function, a hill-climbing search strategy, and the bootstrapping technique, which is able to not only identify a unique causal structure but also quantify the associated uncertainty. Using simulation studies, we demonstrate the power of the proposed learning algorithm by comparing it with alternative Bayesian network methods. For illustration, we consider a dataset from a psychological study of the functional relationships among the symptoms of obsessive-compulsive disorder and depression. Without any prior knowledge, the proposed algorithm reveals some plausible causal relationships. This paper is accompanied by a user-friendly open-source R package `[name hidden]` on CRAN.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

1 Introduction

Survey questionnaires are often used in social, psychological, and behavioral sciences to measure various traits of individuals, which are otherwise hard to assess. For example, Posttraumatic Stress Checklist is often used for measuring post-traumatic stress disorder symptoms, Yale-Brown Obsessive-Compulsive Scale for obsessive compulsive disorder, and Quick Inventory of Depressive Symptomatology for depression, just to name a few.

Various causal inference methods such as the potential outcome framework and structural equation models (SEMs) have been used to infer causal effects.

The potential outcome framework (Neyman, 1923; Rubin, 1974) is a widely used approach to estimate the effects of treatments on outcomes from observational studies. It defines the treatment/causal effect of an experiment unit by contrasting the outcome under the treatment and the outcome under the control. The fundamental challenge of causal inference is that only one of the two potential outcomes can be observed for each experimental unit. Naively using observed outcomes alone to estimate (average) causal effects will be biased due to confounding effects. Properly adjusting for confounders is therefore key to the success of the potential outcome framework.

An SEM refers to a set of stochastic equations describing the statistical causal relationships among observed and latent variables (Jöreskog, 2005; Tarka, 2018). In the psychological field, the latent variables represent latent psychological states or traits, which are believed to exist but difficult to quantify or measure directly, and the observed variables are the "symptoms" or indicators of the latent traits, which, by contrast, can be measured by questionnaires. An SEM is comprised of two components: a measurement/factor model and a structural/path model. The measurement model connects observed variables to latent variables whereas the structural model specifies the relationships among the latent variables, which reflect the causal assumptions made by investigators.

Despite the success of these causal models, alternative causal approaches are called for due to a few prominent limitations. First, both the potential outcome framework and SEMs typically assume the causal relationships among observed/latent variables to be known *a priori*. For example, in the potential outcome framework, one has to know which variables are treatments and which variables are outcomes. However, in many psychological applications, the true causal relationships are unknown, and the inferential results can be quite sensitive with respect to the misspecification

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

of the causal relationships, which could lead to various practical issues such as the Haywood case or the negative variance problem for SEMs (Bentler and Chou, 1987; Kolenikov and Bollen, 2012), and, more seriously, causal effect estimation bias and misinterpretation (Kolenikov, 2011), which are partially responsible for the replicability crisis in psychology and social science and cannot be alleviated by increasing sample size (Vowels, 2021). Second, a latent variable in an SEM often causes multiple symptoms, which implicitly assumes that the symptoms are conditionally independent of each other given the latent variable. However, symptoms can affect each other directly. For example, lack of appetite can cause weight loss and hence they are not independent of each other conditioned on their common cause such as depression. Third, some SEMs can accommodate scenarios where the causal relationships are only partially known. For example, the ordinal SEM (Luo et al., 2021, OSEM) learns the causal structure among latent variables from the data. However, the casual structure is not uniquely identifiable (i.e., multiple causal structures can fit the data equally well), and, therefore, no definitive conclusion can be drawn from such methods.

An alternative class of models for causal analyses is Bayesian networks (Pearl, 1988, BNs), which can overcome the aforementioned limitations of the potential outcome framework and SEMs because BNs typically do not assume the underlying causal structure to be known *a priori*. BNs are a type of probabilistic graphical model that can be used to represent and learn causal relationships of a set of variables in an unbiased, objective, data-driven way. Many fields of science, such as neuroscience (Shen et al., 2020), climate science (Ebert-Uphoff and Deng, 2012), and robotics (Lazkano et al., 2007), have seen rapidly growing enthusiasm for using BNs to discover unknown causal structures. For example, in systems biology, BNs have been shown to successfully recover gene regulatory networks from observational, cross-sectional genomic data without any prior biological knowledge (Friedman et al., 2000; Sachs et al., 2005; Chai et al., 2014; Choi et al., 2020; Zhou et al., 2023; Choi and Ni, 2023).

The potential of BNs to characterize complex causal relationships for survey data in social and behavior sciences has, however, only been demonstrated in a few recent works (McNally et al., 2017; Fried et al., 2017; Bird et al., 2019; Luo et al., 2021; Briganti et al., 2022; Briganti, 2022). Although they already provided compelling evidence that BNs are powerful causal analysis tools, which complement the potential outcome framework and SEMs, their techniques tend to suffer from causal non-identifiability with observational data. For example, Bird et al. (2019) wrote "as

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

distinct causal models can lead to the same patterns, it is not possible to learn all the causal links from observational data." However, since the seminal paper (Shimizu et al., 2006) published in 2006, numerous BNs (e.g., Hoyer et al. 2009; Zhang and Hyvärinen 2009) have been proven to be fully identifiable under various assumptions. Most relevant to the survey questionnaire data is the recent development of ordinal BNs (Ni and Mallick, 2022). They theoretically proved that the causal structure is fully identifiable by exploiting the ordinal nature of categorical data, which had not been thought to be important for causal discovery, and empirically validated it with multiple real datasets such as discretized protein expression data.

Furthermore, non-identifiable BNs such as the commonly used categorical BNs can lead to unintended negative consequences if one is not careful in interpreting the inferred causal networks. Because multiple causal networks can fit the data equally well for non-identifiable BNs, it would be generally incorrect to interpret the causal relationships from a single causal network.

In this paper, we advocate the use of ordinal BNs in social and behavior sciences as a lot of questionnaire data collected are naturally ordinal. We develop a causal structure learning algorithm with bootstrapping, which aims to identify optimal causal structures with finite-sample uncertainty quantification and large-sample guarantee. Using simulation studies, we demonstrate the power of the proposed learning algorithm by comparing it with competing BNs. Subsequently, we apply the ordinal BNs to a dataset of obsessive-compulsive disorder and depression, which reveals some plausible causal relationships without resorting to any prior knowledge. For reproducibility and broad applicability, we make a user-friendly R package `[name hidden]` freely available on CRAN.

Our main contributions are four-fold:

1. We develop a new causal structure learning algorithm with uncertainty quantification.
2. We make a new user-friendly R package available to the scientific community.
3. We introduce a novel application of ordinal BNs to psychological survey data.
4. We provide an asymptotic justification of our method, which guarantees the correctness of the estimated causal graph for a large enough sample size.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

2 Overview of Probabilistic Graphical Models

Let $\boldsymbol{X} = (X_1, \ldots, X_q)$ denote a vector of q random variables. For questionnaire data, X_j represents the available choices of question j, e.g., X_j may take value from "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree" for a 5-point Likert scale question. For convenience, X_j is often coded numerically, e.g., $X_j \in \{1, 2, 3, 4, 5\}$; however, note that the actual number does not have an absolute interpretation but its relative ordering is informative in the sense that $X_j = 1$ is closer to $X_j = 2$ than to $X_j = 5$. To represent the (causal or non-causal) dependencies among a set of random variables, probabilistic graphical models are often used. Let $G = (V, E)$ denote a graph with a set of nodes $V = \{1, \ldots, q\}$ corresponding to the random variables $\boldsymbol{X}$ and a set of edges E representing the dependencies. The type of edges that the edge set E contains dictates the type of dependencies that a graph can represent. We will restrict our discussion to two commonly used types of graphs, undirected graphs and directed acyclic graphs, with a focus on the latter.

Undirected Graphs The edge set E of an undirected graph contains only non-directional edges $X - Y$, which are useful for representing symmetric associations. The presence (absence) of an edge between two variables indicates a statistically significant marginal correlation or partial correlation (the lack thereof). For ordinal variables, the polychoric correlation may be used. Partial correlation is often deemed more appropriate than marginal correlation as partial correlation is a measure of conditional dependence accounting for all the other variables of interest and, therefore, can avoid detecting spurious indirect association from marginal correlation. However, by design, undirected graphs based on marginal or partial correlation cannot be used to represent causal relationships, which are asymmetric and directional.

Directed Acyclic Graphs and Bayesian Networks The edge set E of a directed acyclic graph (DAG) contains only directed edges or arrows $X \rightarrow Y$. In addition, we assume that there is no directed cycle, i.e., one cannot return to the same node by following the arrows. A DAG, by itself, is a pure mathematical object, which needs to be connected to data through probability models. The most well-known probability model of such kind is the *Bayesian network* (BN) proposed by Judea Pearl (Pearl, 1988). A BN is a pair $\mathcal{B} = (G, P)$ where G is a DAG and P is a probability

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

distribution that is linked to the DAG G through the BN factorization,

$$P(\boldsymbol{X}|G) = \prod_{j=1}^{q} P(X_j|\boldsymbol{X}_{\mathrm{pa}(j)}), \tag{1}$$

where $\mathrm{pa}(j) = \{k \in V | k \to j\}$ is the *parent* set of node j and $P(X_j|\boldsymbol{X}_{\mathrm{pa}(j)})$ is the conditional probability distribution of node j given its parents. For categorical variables, $P(X_j|\boldsymbol{X}_{\mathrm{pa}(j)})$ is conditionally multinomial, typically specified by a conditional probability table. For example, if $X_j \in \{\text{Low, Medium, High}\}$ is an employee's pay grade and $\boldsymbol{X}_{\mathrm{pa}(j)} = X_k \in \{$Elementary School, High School, College, Advanced Degrees$\}$ is the employee's education level, then $P(X_j|\boldsymbol{X}_{\mathrm{pa}(j)}) = P(X_j|X_k)$ can be specified by a 3×4 conditional probability table with the first row and second column being the probability $P(X_j = \text{Low}|X_k = \text{High School})$ that an employee is at the low pay grade if his/her highest degree is high school. The BN factorization (1) implies a set of conditional independence assertions, also known as the Markov property, which can be directly read off from G (Lauritzen, 1996). For instance, the probability distribution P must respect the following conditional independence (known as the *local Markov property*): any variable is conditionally independent of its non-descendants given its parents, $X_j \perp \boldsymbol{X}_{\mathrm{nd}(j)}|\boldsymbol{X}_{\mathrm{pa}(j)}$ where $\mathrm{nd}(j) = V\backslash\{j\}\backslash\mathrm{de}(j)$ denotes the set of non-descendants of node j with $\mathrm{de}(j) = \{k \in V | j \to \cdots \to k\}$ being the set of descendants of node j. Importantly, the reverse is also true, i.e., if a distribution P satisfies the local Markov property of a DAG G, it must factorize with respect to G as in (1). For example, for the three-node DAG (h) in Figure 1 where, say, X_3 is an employee's pay grade, X_2 is the employee's education level, and X_1 is the education level of the employee's mother, specifying the joint distribution of (X_1, X_2, X_3) through the conditional distribution $P(X_3|X_2)$ of the employee's pay grade given his/her education level, the conditional distribution $P(X_2|X_1)$ of the employee's education level given his/her mother's education level, and the marginal distribution $P(X_1)$ of the education level of the employee's mother is equivalent to assuming that the employee's pay grade is independent of the education level of the employee's mother given the employee's education level.

Causal DAGs and Causal BNs The arrows of a DAG do not have physical interpretations and, consequently, a BN is merely a probability model that encodes certain conditional independence assertions and factorizes in a certain fashion with respect to its associated DAG. To equip

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

BNs with causal interpretations, we need to first define a causal DAG. A causal DAG is a DAG for which the directed edges are causal. For example, DAG (h) in Figure 1 means that node 1 (e.g., blood pressure) is a direct cause of node 2 (e.g., heart attack), which is in turn a direct cause of node 3 (e.g., death), and node 1 is not a direct cause of node 3. Then we assume a probability distribution P is *causal Markov* with respect to G, i.e., any variable is conditionally independent of its non-effects given its direct causes, $X_j \perp \boldsymbol{X}_{\text{ne}(j)} | \boldsymbol{X}_{\text{dc}(j)}$ where $\text{dc}(j) = \{k \in V | k \to j\}$ is the set of direct causes of j and $\text{ne}(j) = V \backslash \{j\} \backslash \text{eff}(j)$ is the set of non-effects of j with $\text{eff}(j) = \{k \in V | j \to \cdots, k\}$. For instance, in DAG (h), death is conditionally independent of blood pressure given the patient has heart attack (although in real life, abnormal blood pressure can cause death in many ways other than heart attack but the missing arrow between nodes 1 and 3 in DAG (h) excludes alternative causal paths between blood pressure and death in this illustrative example).

Noticing the equivalence in definition between the parents pa(j) and the direct causes $\text{dc}(j)$, and between the non-descendants $\text{nd}(j)$ and the non-effects $\text{ne}(j)$, one can immediately conclude that the probability distribution P must factorize with respect to the causal DAG G as in (1) given the causal Markov assumption. Such a pair of causal DAG G and probability distribution P is called causal BN $\mathcal{B} = (G, P)$. While a non-causal BN says nothing about the true data-generating mechanism, a causal BN does – first, root nodes (i.e., nodes without direct causes) are generated independently from their marginal distributions, and then recursively, a node is generated from the conditional distribution given its direct causes when all of its direct causes have been generated.

A causal BN entails not only the observational distribution P but also distributions subject to various interventions. Formally, let $Q \subset V$ and $I \subset V$ denote the query and intervention sets, and we are interested in calculating the distribution of $\boldsymbol{X}_Q$ if we set $\boldsymbol{X}_I$ to $\boldsymbol{x}_I$ by intervention. Under the Judea Pearl's *do-calculus* paradigm (Pearl, 1988), that amounts to finding the interventional probability distribution $P(\boldsymbol{X}_Q | \text{do}(\boldsymbol{X}_I = \boldsymbol{x}_I), G)$ where the do-operator $\text{do}(\boldsymbol{X}_I = \boldsymbol{x}_I)$ highlights the fact that $\boldsymbol{X}_I$ is set to $\boldsymbol{x}_I$ by intervention, not observed to be $\boldsymbol{x}_I$. This interventional probability distribution is generally not equal to the conditional distribution $P(\boldsymbol{X}_Q | \boldsymbol{X}_I = \boldsymbol{x}_I, G)$ induced from the joint observational distribution $P(\boldsymbol{X} | G)$. In fact, $P(\boldsymbol{X}_Q | \text{do}(\boldsymbol{X}_I = \boldsymbol{x}_I), G) = P(\boldsymbol{X}_Q | \boldsymbol{X}_I = \boldsymbol{x}_I, G_I)$ where G_I is a "mutilated" version of G with all incoming arrows to I removed. Intuitively, when there is no intervention, the value of $\boldsymbol{X}_I$ is influenced by its direct causes whereas when $\boldsymbol{X}_I$

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

is set to a certain value by intervention, such a value only depends on the intervention[1]. Therefore, in the presence of intervention, the incoming arrows to $\boldsymbol{X}_I$ should be removed to reflect the fact that $\boldsymbol{X}_I$ is no longer influenced by its natural causes. Take DAG (t) in Figure 1 as an example. From basic probability theory, the conditional distribution of X_3 given that X_2 is observed to be x_2 is $P(X_3|X_2 = x_2) = \sum_{X_1} P(X_3|X_1, X_2 = x_2)P(X_1|X_2 = x_2)$. However, if X_2 is not naturally observed to be x_2 but instead we set its value to x_2 by intervention, the mutilated version of DAG (t) is given by DAG (o) where the arrow from node 1 to node 2 is removed, and the interventional distribution is

$$P(X_3|\mathrm{do}(X_2 = x_2)) = \sum_{X_1} P(X_3|X_1, X_2 = x_2)P(X_1), \tag{2}$$

because X_1 and X_2 are marginally independent in DAG (o). Being able to derive various interventional distributions using causal BNs is crucial to social and behavior sciences as it does not require real-world interventions, which may be expensive, unethical, or impossible to carry out. For instance, let X_1 denote age, X_2 cortical thickness, and X_3 intelligence in the study of the relationship between brain structure and intelligence (Shaw et al., 2006). Suppose the causal relationships of X_1, X_2, and X_3 are represented by DAG (t). To identify the average causal effect of cortical thickness on intelligence, i.e., $ACE = E(X_3|\mathrm{do}(X_2 = 1)) - E(X_3|\mathrm{do}(X_2 = 0))$ (say, $X_2 = 0$ and 1 respectively represent thin and thick cortex), the gold standard would be to intervene on cortical thickness, which, however, cannot be done. Causal BNs enable such causal effect estimation without carrying out the actual intervention via (2); notice that the right-hand side of (2) does not have the do-operator and hence can be calculated based on observational probability distribution P alone.

Learning of Causal DAGs and BNs The preceding paragraphs concern the problem of representation, i.e., given a causal DAG, how one can represent the (stochastic) data-generating mechanism using a probability model. The remaining question is whether one can learn the unknown structure of DAG G given a sample generated from the probability model, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim P(\boldsymbol{X}|G)$. One intuitive approach is to test for independence. For instance, consider three-node

[1] Such intervention is also known as the hard intervention in causal discovery literature.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

DAGs (there are 25 in total shown in Figure 1), and suppose DAG (h) in Figure 1 is the true data-generating DAG and a large number of observations are available. Assuming causal faithfulness[2], we sequentially test for independence (i) between X_1 and X_2, which comes to be dependent $X_1 \not\perp X_2$ and hence eliminates all DAGs in the blue boxes as they all encode $X_1 \perp X_2$, (ii) between X_2 and X_3, which comes to be dependent $X_2 \not\perp X_3$ and hence eliminates all DAGs in the green boxes as they all encode $X_2 \perp X_3$, (iii) between X_1 and X_3, which eliminates DAG (k) in the yellow box, and (iv) finally between X_1 and X_3 given X_2, which comes to be independent $X_1 \perp X_3 | X_2$ and hence eliminates all DAGs in the purple boxes as they assert dependence $X_1 \not\perp X_3 | X_2$. This example demonstrates that just by applying independence tests on observed data, one can narrow down from 25 possible DAGs to just three DAGs (h-j) that are plausible data-generating mechanisms. This type of approach is called the constraint-based approach. The PC algorithm (Spirtes et al., 2000) is perhaps the most well-known one. However, there are obvious drawbacks of constraint-based approaches: apart from the additional assumption of faithfulness and conditional independence tests generally lacking statistical power, most prominently, they generally can only identify an equivalence class of DAGs, all of which encode exactly the same conditional independence relationships; such DAGs and corresponding BNs are said to be *Markov equivalent* and the equivalence classes are called Markov equivalence classes. In the three-node example, DAGs (h)-(j) have the same set of conditional independencies, i.e., $X_1 \perp X_3 | X_2$ and none other. Therefore, one cannot further narrow it down to the true data-generating DAG (h) even with an infinite sample. This is clearly an unsatisfactory property of constraint-based approaches as DAGs (h)-(j) have very different causal interpretations from each other.

Another major class of causal DAG learning approaches is score-based where one would assign a score to each DAG and search for highly scored DAGs. Often, the score is based on some probability model and depends on the likelihood. For example, the Bayesian information criterion (BIC) is widely used,

$$\mathrm{BIC}(G|\boldsymbol{x}) = -2\sum_{i=1}^{n} \log \widehat{P}(\boldsymbol{x}_i|G) + K\log(n), \tag{3}$$

[2] A distribution P is said to be faithful to a causal DAG G if all conditional independence relationships of P are encoded in G through its Markov property.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

where K is the number of model parameters and $\widehat{P}(\boldsymbol{x}_i|G)$ is the joint distribution (1) evaluated at $\boldsymbol{x}_i$ given the maximum likelihood estimate of model parameters. BIC balances between the goodness-of-fit of the causal BN to the observed data and the complexity of the model. For categorical data, $P(\boldsymbol{x}_i|G)$ is specified by conditional probability tables as mentioned earlier. Unfortunately, it can be shown that DAGs (h)-(j) are score-equivalent and are, thus, still indistinguishable from each other just like constraint-based methods. We illustrate it with DAGs (h)&(i). Suppose again DAG (h) is the true data-generating DAG and the corresponding conditional probability tables are given in Figure 2(a). These conditional probability tables determine the joint probability distribution of (X_1, X_2, X_3), for example, $P(X_1 = 1, X_2 = 2, X_3 = 3) = P(X_1 = 1)P(X_2 = 2|X_1 = 1)P(X_3 = 3|X_2 = 2) = 0.25 \times 0.19 \times 0.50 = 0.024$. However, the joint distribution $P(X_1, X_2, X_3)$ can be factorized in a few other ways that are also compatible with the conditional independence relationship $X_1 \perp X_3|X_2$ encoded in DAG (h). For example, it can be factorized with respect to DAG (i) in Figure 1 of which the conditional probability tables are shown in Figure 2(b). Consequently, DAGs (h)&(i) have the same BIC score because they have the same maximized likelihood and the same model complexity, and hence cannot be distinguished from each other. This also applies to DAG (j) in Figure 1. More generally, Markov equivalent categorical BNs (cBNs) are (BIC) score-equivalent. Therefore, score-based cBNs cannot differentiate DAGs that are indistinguishable by the constraint-based methods. More discussion of categorical causal BNs is provided in Section 6.

We remark that many existing DAG learning algorithms would return a single DAG even though the underlying causal model is not fully identifiable (i.e., there exist equivalent DAGs). Practitioners should be aware that the returned DAG is generally an arbitrary choice from its equivalence class and there may, in fact most likely, exist many other DAGs that fit the data equally well and have very different causal implications. Therefore, we recommend the use of identifiable causal models (e.g., the ordinal BN in the next section) whenever possible.

3 Ordinal Bayesian Networks

Probability Model Since a lot of questionnaire data in social and behavior sciences are ordinal, we propose the use of ordinal BNs (Ni and Mallick, 2022, oBN) to resolve the indeterminacy

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

of Markov/score-equivalent BNs. Ni and Mallick (2022) theoretically studied the causal identifiability of oBN and showcased its strength in constructing biological networks from observational, discretized protein expression data. oBN can potentially have great utility for discovering causality in questionnaire data. Let $X_j \in \{1, \ldots, L_j\}$ have L_j categories for $j = 1, \ldots, q$. Each conditional distribution $P\left(X_j|\boldsymbol{X}_{\text{pa}(j)}\right)$ of (1) takes the form of an ordinal regression model of which the cumulative distribution is given by, for $\ell = 1, \ldots, L_j$,

$$P(X_j \leqslant \ell|\boldsymbol{X}_{\text{pa(j)}}) = F\left(\gamma_{j\ell} - \sum_{k \in \text{pa(j)}} \beta_{jkX_k} - \alpha_j\right),$$

where F is a link function such as probit and logistic, α_j is an intercept, β_{jkX_k} is a generic notation of $\beta_{jk1}, \ldots, \beta_{jkL_k}$ for $X_k = 1, \ldots, L_k$, and $\gamma_{j1} < \cdots < \gamma_{jL_j} = \infty$ are a set of thresholds. We set $\gamma_{j1} = \beta_{jk1} = 0$ for ordinal regression parameter identifiability (Agresti, 2003). The implied conditional probability distribution is given by,

$$\begin{aligned} P(X_j = \ell|\boldsymbol{X}_{\text{pa(j)}} = \boldsymbol{x}_{\text{pa(j)}}) = F(\gamma_{j\ell} - \sum_{k \in \text{pa(j)}} \beta_{jkx_k} - \alpha_j) \\ -F(\gamma_{j,\ell-1} - \sum_{k \in \text{pa(j)}} \beta_{jkx_k} - \alpha_j), \end{aligned}$$

for $\ell = 1, \ldots, L_j$ and $x_k \in \{1, \ldots, L_k\}$ for $k \in \text{pa(j)}$.

To illustrate the identifiability of oBN, consider again the example in Figure 2. Let G_1 and G_2 be the DAGs in 2(a) and 2(b), respectively. While a cBN can be factorized in either direction, by exploiting the ordinal nature of categorical data, an oBN does not admit such equivalent factorization. In fact, the conditional probability tables in Figure 2(a) are those under the oBN with DAG G_1 and the following parameter values,

$\gamma_{12} = \alpha_1 = 0.67,$

$\gamma_{22} = 0.5, \alpha_2 = 0.5, \beta_{212} = -0.5, \beta_{213} = 0.75,$

$\gamma_{32} = 0.5, \alpha_3 = -1, \beta_{322} = 0.5, \beta_{323} = -0.75,$

whereas there are no parameter values of the oBN with DAG G_2 such that the implied observational distribution $P(X_1, X_2, X_3|G_2)$ is compatible with the conditional probability tables in Figure 2(b). In other words, G_1 and G_2 are not score-equivalent as they have different likelihood functions.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Statistical Inference We now develop a score-based DAG learning algorithm, which aims to identify the best-scored DAG with uncertainty quantification. We score each DAG G with the BIC (3) where the maximum likelihood estimate is obtained by gradient ascent, and the number of model parameters is $K = \sum_{j \in V}(L_j - 1 + \sum_{k \in \mathrm{pa}(j)}(L_k - 1))$. To search for the best-scored DAG, we use an iterative hill-climbing algorithm. We start from some initial DAG. At each iteration, we score all the DAGs that are reachable from the current graph by a single edge addition, removal, or reversal. We replace the current DAG by the DAG with the largest improvement (i.e., the largest decrease in BIC). We claim the convergence of the algorithm when the BIC can no longer be improved. The hill-climbing algorithm is summarized in Algorithm 1. Since BIC always improves at each iteration by design, the algorithm is guaranteed to find a local optimum.

However, there are two drawbacks of Algorithm 1. First, the local optimum may not be the global optimum due to the greedy nature of the hill-climbing algorithm. Therefore, we suggest repeat the hill-climbing algorithm several times with random initial DAGs and pick the DAG with the smallest BIC as shown in Algorithm 2.

Second, Algorithm 1 or 2 only provides a point estimate of DAG G without uncertainty quantification. To assess the uncertainty, we propose to use the bootstrapping technique (Efron, 1992; Friedman et al., 1999). Specifically, we first create a number B of bootstrap samples by sampling without replacement from the original data $\boldsymbol{x}$. Then, we apply Algorithm 2 to each bootstrap sample. Finally, we compute the average adjacency matrix of the estimated DAG from each bootstrap sample. An adjacency matrix $\boldsymbol{A} = [A_{jk}]$ of DAG G is a binary matrix such that $A_{jk} = 1$ if $k \to j \in E$ and $A_{jk} = 0$ otherwise. Therefore, the average adjacency matrix, denoted by $\boldsymbol{P} = [P_{jk}]$, can be interpreted as an approximate edge inclusion probability of $k \to j$. A value of $P_{jk} := \frac{1}{B}\sum_{b=1}^{B} A_{jk}^{(b)}$ close to 0 or 1 indicates greater confidence of the absence or presence of $k \to j$ than a value close to 0.5 where the superscript (b) indexes the bootstrap samples. The hill-climbing with multiple initial DAGs and bootstrapping is described in Algorithm 3 and implemented in the R package `[name hidden]` freely available on CRAN.

Large Sample Property Now, we ask if we can correctly identify the data-generating DAG when the sample size is large enough. Let G_1 denote the true data-generating DAG with model parameters $\theta_1^\star$ (i.e., α, β, γ's in oBN). Let G_2 denote any other DAG in the same Markov equivalence

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Algorithm 1 Hill-Climbing: $[G, \mathrm{BIC}] = \mathrm{HC}(\boldsymbol{x}, G_0)$

```
Input: data $\boldsymbol{x}$, initial DAG $G_0$
Set $G = G_0$, compute $\mathrm{BIC}(G|\boldsymbol{x})$, and set $\mathrm{BIC}_\star$=$\mathrm{BIC}(G|\boldsymbol{x})$
repeat
  Initialize $Improvement = false$
  for all graphs $G'$ reachable from $G$ do
    Compute $\mathrm{BIC}(G'|\boldsymbol{x})$
    if $\mathrm{BIC}(G'|\boldsymbol{x}) < \mathrm{BIC}_\star$ then
      Set $G = G'$ and $\mathrm{BIC}_\star$=$\mathrm{BIC}(G'|\boldsymbol{x})$
      Set $Improvement = true$
    end if
  end for
until $Improvement$ is $false$
Output: estimated DAG $G$ and its BIC = $\mathrm{BIC}(G|\boldsymbol{x})$
```

Algorithm 2 Hill-Climbing with Multiple Initial DAGs: $G = \text{HC-x}(\boldsymbol{x}, R)$

```
Input: data $\boldsymbol{x}$, number of initial DAGs $R$
for $r = 1, \dots, R$ do
  Randomly generate a DAG $G_{0r}$
  $[G_r, \mathrm{BIC}_r] = \mathrm{HC}(\boldsymbol{x}, G_{0r})$
end for
Set $G = G_r$ with $r = \arg\min_{r'} \mathrm{BIC}_{r'}$
Output: estimated $G$
```

Algorithm 3 Hill-Climbing with Multiple Initial DAGs and Bootstrapping: $P = \text{HC-x-Boot}(\boldsymbol{x}, R, B)$

```
Input: data $\boldsymbol{x}$, number of initial DAGs $R$, number of bootstrap samples $B$
for $b = 1, \dots, B$ do
  Generate a bootstrap sample $\boldsymbol{x}^b$ of $\boldsymbol{x}$ by sampling with replacement
  $G^b = \text{HC-x}(\boldsymbol{x}^b, R)$
end for
Set $\boldsymbol{P} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{A}^b$ where $\boldsymbol{A}^b$ is the adjacency matrix of $G^b$
Output: estimated edge inclusion probability matrix $\boldsymbol{P}$
```

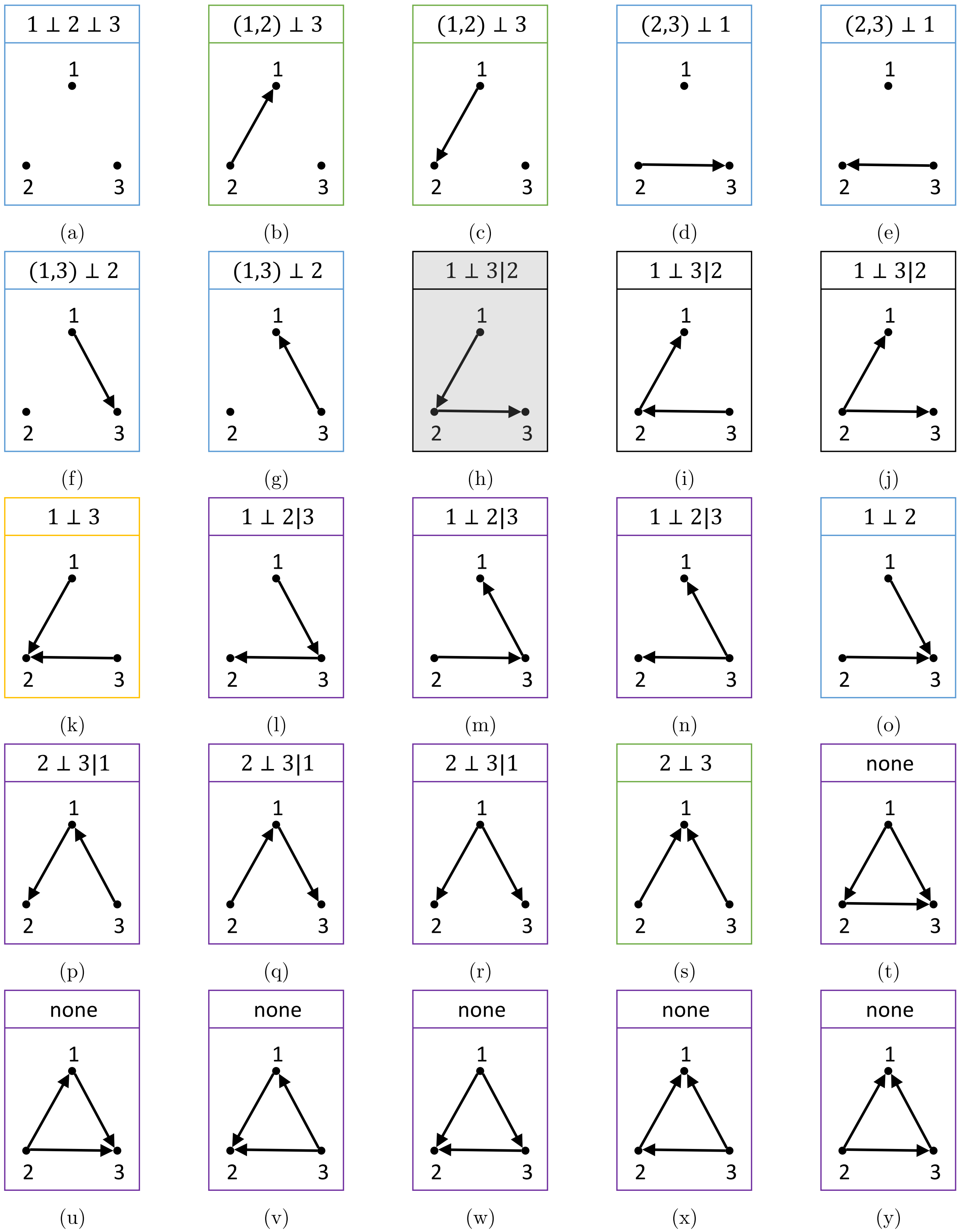
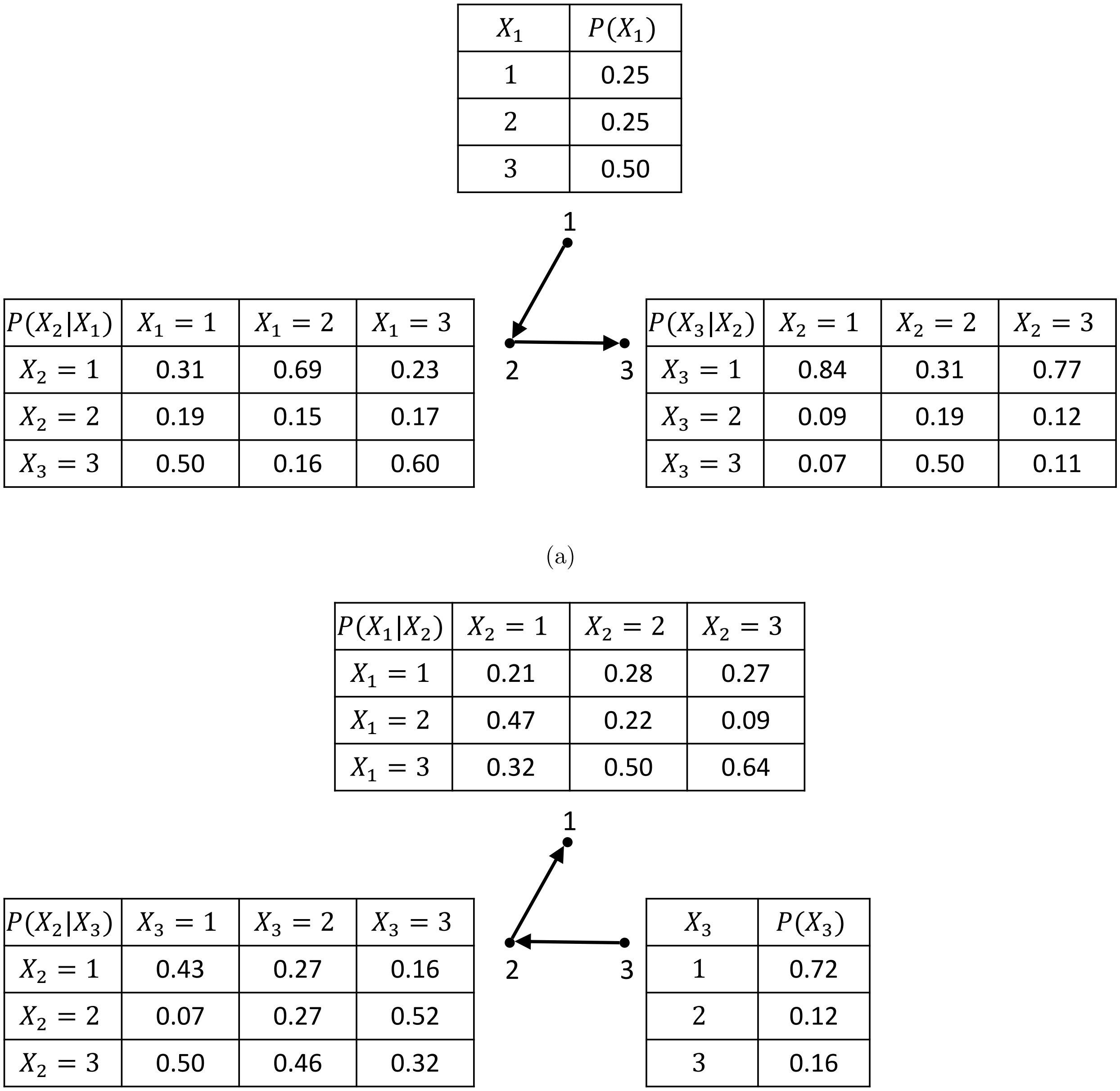


Figure 1: All possible three-node DAGs. The conditional independence assertion encoded by each graph is shown at the top of each DAG.



X_1	$P(X_1)$
1	0.25
2	0.25
3	0.50

$P(X_2 \mid X_1)$	$X_1 = 1$	$X_1 = 2$	$X_1 = 3$
$X_2 = 1$	0.31	0.69	0.23
$X_2 = 2$	0.19	0.15	0.17
$X_3 = 3$	0.50	0.16	0.60

$P(X_3 \mid X_2)$	$X_2 = 1$	$X_2 = 2$	$X_2 = 3$
$X_3 = 1$	0.84	0.31	0.77
$X_3 = 2$	0.09	0.19	0.12
$X_3 = 3$	0.07	0.50	0.11

(a)

$P(X_1 \mid X_2)$	$X_2 = 1$	$X_2 = 2$	$X_2 = 3$
$X_1 = 1$	0.21	0.28	0.27
$X_1 = 2$	0.47	0.22	0.09
$X_1 = 3$	0.32	0.50	0.64

$P(X_2 \mid X_3)$	$X_3 = 1$	$X_3 = 2$	$X_3 = 3$
$X_2 = 1$	0.43	0.27	0.16
$X_2 = 2$	0.07	0.27	0.52
$X_2 = 3$	0.50	0.46	0.32

X_3	$P(X_3)$
1	0.72
2	0.12
3	0.16

(b)

Figure 2: Conditional probability tables from two Markov equivalent BNs.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

class with G_1. Let $\theta_2^\dagger$ denote its pseudo-true parameter, i.e.,

$$\theta_2^\dagger = \arg\max_\theta \sum_{\boldsymbol{X}} P(\boldsymbol{X}|G_1, \theta_1^\star) \log P(\boldsymbol{X}|G_2, \theta).$$

We show that the BIC of G_1 is asymptotically lower than that of G_2. Let $\ell_n(\theta|G) = \sum_{i=1}^n \log P(\boldsymbol{X}_i = \boldsymbol{x}_i|G, \theta)$ denote the log-likelihood under DAG G and let $\widehat{\theta}_1^{(n)}$ and $\widehat{\theta}_2^{(n)}$ denote the maximum likelihood estimators, which are consistent under some mild regularity conditions (Fahrmeir and Kaufmann, 1985), i.e., $\widehat{\theta}_1^{(n)} \xrightarrow{p} \theta_1^\star$ and $\widehat{\theta}_2^{(n)} \xrightarrow{p} \theta_2^\dagger$ as $n \to \infty$.

Take the Taylor expansion of $\ell_n(\theta_1^\star|G_1)$ at $\widehat{\theta}_1^{(n)}$,

$$\begin{aligned}\ell_n(\theta_1^\star|G_1) &= \ell_n(\widehat{\theta}_1^{(n)}|G_1) + (\theta_1^\star - \widehat{\theta}_1^{(n)})^\top \left.\frac{\partial \ell_n(\theta|G_1)}{\partial\theta}\right|_{\widehat{\theta}_1^{(n)}} + \frac{1}{2}(\theta_1^\star - \widehat{\theta}_1^{(n)})^\top \left.\frac{\partial^2 \ell_n(\theta|G_1)}{\partial\theta^2}\right|_{\xi^{(n)}} (\theta_1^\star - \widehat{\theta}_1^{(n)}),\\ &= \ell_n(\widehat{\theta}_1^{(n)}|G_1) + \frac{1}{2}(\theta_1^\star - \widehat{\theta}_1^{(n)})^\top \left.\frac{\partial^2 \ell_n(\theta|G_1)}{\partial\theta^2}\right|_{\xi^{(n)}} (\theta_1^\star - \widehat{\theta}_1^{(n)})\end{aligned}$$

where $\xi^{(n)} = \alpha\theta_1^\star + (1-\alpha)\widehat{\theta}_1^{(n)}$ with $\alpha \in [0, 1]$. Since $\widehat{\theta}_1^{(n)} \xrightarrow{p} \theta_1^\star$, we have $\frac{1}{n}\ell_n(\widehat{\theta}_1^{(n)}|G_1) - \frac{1}{n}\ell_n(\theta_1^\star|G_1) \xrightarrow{p} 0$. By the law of large numbers,

$$\frac{1}{n}\ell_n(\theta_1^\star|G_1) \xrightarrow{p} E[\log P(\boldsymbol{X}|G_1, \theta_1^\star)],$$

where the expectation is taken over $\boldsymbol{X}$ with respect to its true data-generating distribution $P(\boldsymbol{X}|G_1, \theta_1^\star)$. Therefore,

$$\frac{1}{n}\ell_n(\widehat{\theta}_1^{(n)}|G_1) \xrightarrow{p} E\left[\log P(\boldsymbol{X}|G_1, \theta_1^\star)\right].$$

By a similar argument, we have

$$\frac{1}{n}\ell_n(\widehat{\theta}_2^{(n)}|G_2) \xrightarrow{p} E\left[\log P(\boldsymbol{X}|G_2, \theta_2^\dagger)\right].$$

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Hence,

$$\frac{1}{n}\ell_n(\widehat{\theta}_1^{(n)}|G_1) - \frac{1}{n}\ell_n(\widehat{\theta}_2^{(n)}|G_2) \xrightarrow{p} E\left[\log\frac{P(\boldsymbol{X}|G_1,\theta_1^\star)}{P(\boldsymbol{X}|G_2,\theta_2^\dagger)}\right]$$

$$= \mathrm{KL}(P(\boldsymbol{X}|G_1,\theta_1^\star)||P(\boldsymbol{X}|G_2,\theta_2^\dagger)) > 0,$$

where $\mathrm{KL}(\cdot||\cdot)$ is the Kullback–Leibler divergence, which is nonnegative and is zero only when $P(\boldsymbol{X}|G_1,\theta_1^\star) \equiv P(\boldsymbol{X}|G_2,\theta_2^\dagger)$, which is impossible due to the causal identifiability result (Ni and Mallick, 2022). Consequently,

$$\ell_n(\widehat{\theta}_1^{(n)}|G_1) - \ell_n(\widehat{\theta}_2^{(n)}|G_2) \xrightarrow{p} \infty.$$

Because two Markov equivalent DAGs must have the same skeleton (Verma and Pearl, 2022) and hence the same model complexity, we have

$$\mathrm{BIC}(G_1|\boldsymbol{x}) - \mathrm{BIC}(G_2|\boldsymbol{x}) \xrightarrow{p} -\infty.$$

4 Simulation Studies

We assessed the empirical performance of oBN in recovering unknown DAG structure using simulations where the ground truth is known. We simulated data with $q = 10$ categorical variables each with 5 ordinal categories resembling the 5-point Likert scale questions. The true DAG was generated randomly using the function “randomDAG” in R package `pcalg` with connecting probability 0.2 (Figure 3a). Its Markov equivalence class, represented by the completed partially directed acyclic graph (CPDAG), is shown in Figure 3b where the bidirected edges are edges that can be oriented in either direction without changing its conditional independence relationships. Given the true DAG, the model parameters $\beta_{jk\ell}$'s and α_j's were independently generated from a centered normal distribution with variance σ^2. We considered 14 scenarios. The first 7 scenarios fixed sample size at $n = 500$ and varied the signal strength $\sigma = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2$, which covered low to strong levels of signals. The other 7 scenarios fixed the signal strength at $\sigma = 2$ and varied the sample size $n = 500, 1000, 2000, 4000, 8000, 16000, 32000$.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Algorithm 1, implemented in R package [name hidden], was applied to each simulated dataset. For comparison, we also ran the PC algorithm and the (nominal) cBN. For the PC algorithm, we used a more recent version (Colombo et al., 2014) implemented as "pc.stable()" in the R package bnlearn with the Jonckheere-Terpstra test designed for ordinal data and the type I error controlled at 1%. For the cBN, we used the BIC scoring criterion and the hill-climbing search algorithm with 10 random starts. cBN is also available in the package bnlearn implemented as "hc()".

As an error measure, we computed the structural hamming distance (SHD) between the estimated graph and the simulation true DAG, which is the number of edge additions, deletions, or reversals required to transform one graph to the other. Note that since cBN and PC can only identify CPDAG (i.e., equivalence classes), the smallest SHD that they can achieve is 4 (the number of bidirected edges in Figure 3b). This error cannot be further reduced for cBN and PC even with an infinite amount of data.

The SHD averaged over 50 repeat simulations are reported as functions of signal strength σ (Figure 4a) and sample size n (Figure 4b). Several conclusions can be made. First, because oBN is a fully identifiable model, its SHD quickly approached 0 as signal became stronger; such trend was not observed for cBN and PC. Second, oBN consistently outperformed cBN and PC across all signal levels and sample sizes, which stresses the importance of accounting for the ordinal nature of questionnaire data for causal discovery, which had been overlooked in the literature. Third, when the sample size was moderate ($n = 500$), the performance of cBN and PC did not improve as the signal strength increased whereas when the signal was strong $\sigma = 2$, their performance improved as sample size grew. Eventually, they might reach the irreducible error (SHD=4 in this example) but that would require a huge amount of data and they cannot do better even with an infinite amount of data. The size of the irreducible error depends on the true data-generating DAG, which could be as large as the total number of edges, which is super-exponential in the number q of variables. Fourth, for a large enough sample, oBN perfectly recovered the true DAG, empirically verifying our asymptotic theory.

In summary, our simulation studies suggest that it is advantageous to exploit the ordinal nature of survey questionnaire data for causal discovery. Considering oBN is conceptually similar to cBN but with better theoretical and empirical properties for causal discovery, we hope to see a wider adoption of oBNs in social and behavioral sciences. In the following, we conducted additional

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

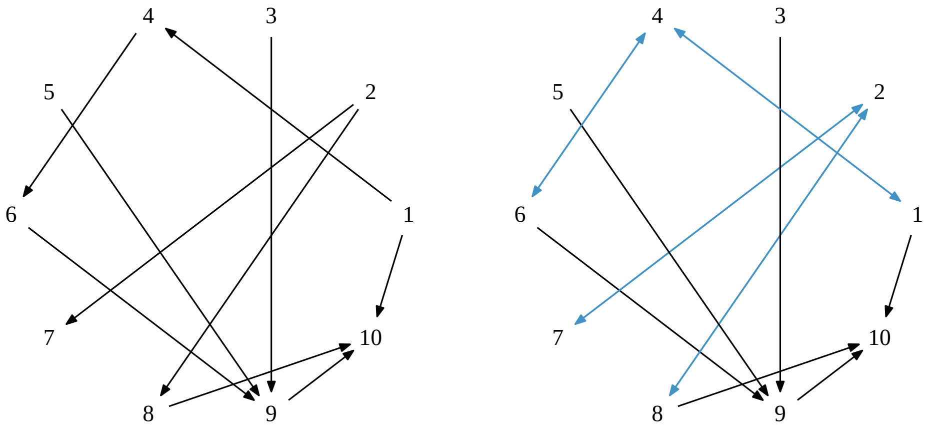
simulations to test the scalability and sensitivity of our method.

Scalability. We varied the number of variables $q = 10, 20, 30, 40, 50$ while keeping the sample size at $n = 500$ and the signal strength at $\sigma = 2$. The data generation process was the same as before. The (normalized) SHD and the CPU time on a 2.9 GHz 6-Core Intel Core i9 CPU averaged over 50 repeat simulations are reported as functions of q (Figure 5). As expected, the performance for all methods deteriorated as q increased but the proposed oBN still outperformed both cBN and PC, and the computation of oBN scaled reasonably well with q.

Sensitivity to link functions. We fitted the proposed model to the data that we simulated earlier (with sample size $n = 500$, number of variables $q = 10$, signal strength $\sigma = 1$, and the probit link function) using probit, logistic, negative log-log, and complementary log-log link functions. The average SHD between the estimated and true graphs based on 50 repeat simulations is reported in Table 1, which shows our model is relatively robust – the SHDs are well within two standard errors from each other.

Probit	Logistic	Negative log-log	Complementary log-log
0.26 (0.06)	0.30 (0.07)	0.32 (0.07)	0.28 (0.06)

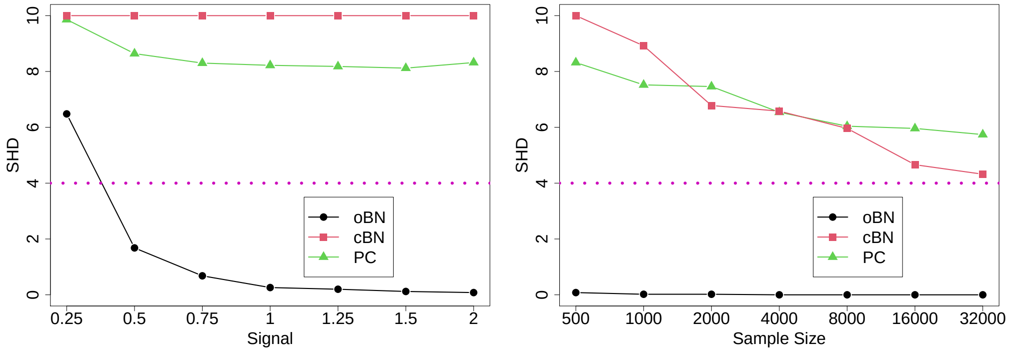
Table 1: Sensitivity to the choice of link functions. The average (standard error) SHD is reported.



(a) Simulation true DAG (b) Simulation true CPDAG

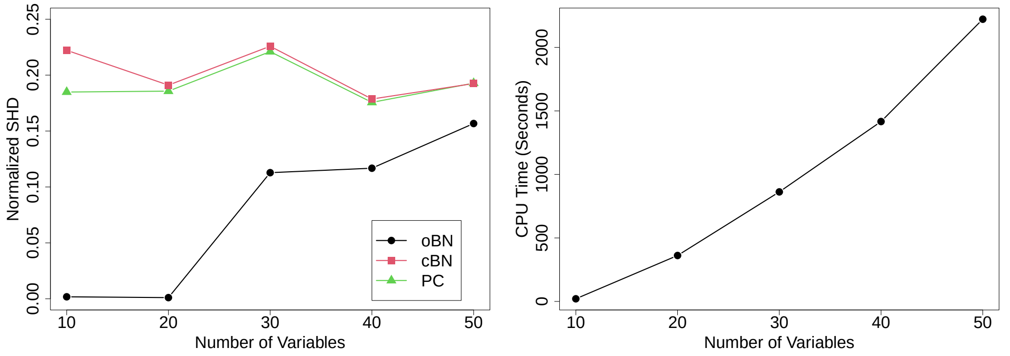
Figure 3: Simulation true (a) DAG and (b) CPDAG. The (blue) bidirected edges in (b) are edges that can be oriented in either direction in the Markov equivalence class represented by the CPDAG.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press



(a) (b)

Figure 4: Simulated survey data with ten 5-point Likert scale questions. Panel (a): Sample size $n = 500$ and signal strength varies from 0.25 to 2. Panel (b): Signal strength $\sigma = 2$ and sample size varies from 500 to 32000. In both panels, dotted lines indicate the irreducible error (SHD=4) for an oracle cBN.



(a) (b)

Figure 5: Simulated survey data with varying number of 5-point Likert scale questions $q = 10, 20, 30, 40, 50$. The sample size is fixed at $n = 500$ and the signal strength is fixed at $\sigma = 2$. Left panel: The SHD is normalized by dividing the raw SHD by the total number of edges in a complete DAG (i.e., $\frac{q(q-1)}{2}$). Right panel: CPU time to convergence of oBN in seconds tested on a 2.9 GHz 6-Core Intel Core i9 CPU.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

5 Demonstration: OCD-Depression Data Analyses

To further demonstrate oBN, we analyzed the dataset from a psychological study of the functional relationships between the symptoms of obsessive-compulsive disorder (OCD) and depression (McNally et al., 2017). The dataset consists of $n = 408$ participants' responses to 10 five-point questions from the Yale-Brown Obsessive-Compulsive Scale via Self-Report (Steketee et al., 1996) measuring the OCD symptoms and 16 four-point questions from the Quick Inventory of Depressive Symptomatology via Self-Report (Rush et al., 2003, QIDS-SR) measuring the depression symptoms. Following Luo et al. (2021), we merged the questions about "decreased appetite" and "increased appetite", and the questions about "weight loss" and "weight gain" in QIDS-SR since they measure the same depression symptoms. The resulting number of ordinal variables is $q = 24$.

Algorithm 3 was applied to the dataset with $R = 10$ random initial DAGs and $B = 500$ bootstrap samples. For an illustration of the guaranteed convergence, we plot the BIC as a function of iteration in one run of our algorithm in Figure 6, which converged at the 37th iteration.

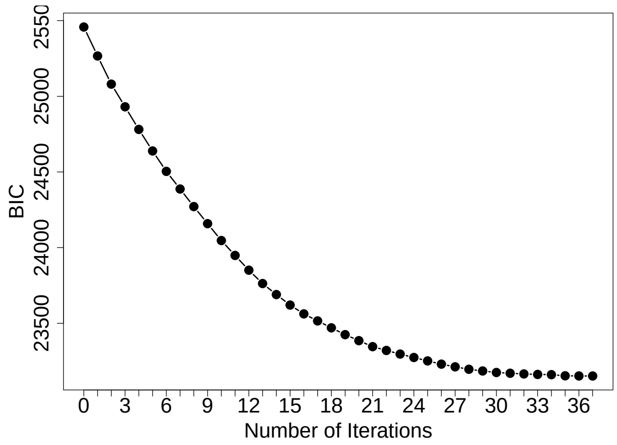


Figure 6: BIC as a function of iteration on the OCD-Depression data.

We also explored a more scalable version of oBN, a two-step hybrid algorithm. Particularly, we first ran the PC algorithm to obtain a CPDAG, and then, for any pair of nodes with an undetermined edge (i.e., the edge can be oriented in either direction), we ran the bivariate version of oBN to determine its direction; the pseudocode is presented in Algorithm 4.

In Figure 7, we plot the estimated DAG from oBN with edge width proportional to the inclusion probability; also see the list of all the significant edges (i.e., $P_{ij} > 0.5$) ranked by their inclusion probabilities in Table 2. For brevity, we have adopted the same abbreviation of the symptoms as

Algorithm 4 PC+oBN: $G = \text{PC-oBN}(\boldsymbol{x})$

```
Input: data $\boldsymbol{x}$
Run PC algorithm on $\boldsymbol{x}$ and obtain a CPDAG $\widetilde{G}$
for each pair $(j, k)$ with an undetermined edge in $\widetilde{G}$ do
  Run Algorithm 1 on the subset of data restricted to $(X_j, X_k)$ and orient the edge according to the output
end for
Output: estimated DAG $G$
```

in McNally et al. (2017). Comparing to the estimated network by PC+oBN (Figure 8), all the undetermined edges from the PC algorithm (Figure 9) were oriented consistently between oBN and PC+oBN. For comparison, we also applied the PC algorithm with the Jonckheere-Terpstra test, the cBN with BIC and hill-climbing, and the ordinal structural equation model (Luo et al., 2021, OSEM) with the caveat that we do not know the underlying true causal relationships. Their results are reported in Figures 9-11. Some interesting observations can be made.

Common to all the methods, the symptoms of OCD and the symptoms of depression were found to be largely separated meaning that most of the symptoms of OCD do not directly cause most of the symptoms of depression, and vice versa. This is perhaps not surprising given that OCD and depression are different psychological disorders. oBN, cBN, and OSEM did simultaneously find one bridge causal link between OCD (*obdistress*) and depression (*sad*). The existence of a bridge causal link is plausible because many studies have suggested that more than one third of OCD patients have concurrent depression (Nestadt et al., 2001; Abramowitz, 2004; Hong et al., 2004). cBN and OSEM, due to their non-identifiability, could not determine the direction of that casual link whereas oBN identified it to be *obdistress*→*sad*. This again seems to agree with existing studies that OCD symptoms often precede depression in individuals who suffer from both disorders (Anholt et al., 2011; Meyer et al., 2014; Zandberg et al., 2015). Our finding suggests that controlling distress caused by obsession may help alleviate or prevent depression symptoms.

Within the symptoms of OCD, on the one hand, the links among obsessive symptoms and the links among compulsive symptoms are the links that tend to have the highest probabilities, e.g., *obinterfer*→*obdistress* (0.846), *obdistress*→*obtime* (0.98), and *compinterf*→*comptime* (0.79). This matches the hypothesized two dimensions of obsession and compulsion in theoretical models (de Wildt et al., 2005). On the other hand, there also exist significant links from compulsion symptoms to their obsession counterparts, including *compinterf→obinterfer*, *comptime→obtime*, and *compresis→obresist*, which are qualitatively consistent with previous network analyses (Carbonella, 2018; Cervin et al., 2020), although their network models are undirected/non-causal. According to our results, one can potentially suppress obsession symptoms by suppressing the corresponding compulsion symptoms but not vice versa.

Within the symptoms of depression, the link between *sad* and *suicide* was found by all the methods but only oBN and PC+oBN were able to determine its direction *sad→suicide* (inclusion probability=0.914 for oBN). It is well-known that persistent feeling of sadness is a major risk factor for suicide (Angst et al., 1999; Brådvik, 2018). Another expected link was *appetite→weight*, of which the direction was again only identified by oBN and PC+oBN.

In summary, although the ground truth is not available, our data analyses, in our opinion, support the usefulness of oBN for generating plausible psychological hypotheses in practice.

6 Discussion

We have demonstrated the functionality of oBNs as a useful alternative to the potential outcome framework and SEMs for analyzing survey questionnaire data. oBNs can provide an unbiased causal view of a complex system without any prior structural knowledge. Moreover, unlike other existing BNs for categorical data, oBNs are fully identifiable with observational data alone.

Note that both cBN and oBN utilize the BIC score. Therefore, the difference between cBN and oBN lies in the likelihood function and the model complexity. For oBN, the likelihood function is specified by ordinal regression whereas for the cBN, it is specified by multinomial distribution. The model complexity is different between the two methods, $K_{\text{oBN}} = \sum_{j\in V}(L_j - 1 + \sum_{k\in \text{pa}(j)}(L_k - 1))$ for oBN and $K_{\text{cBN}} = \sum_{j\in V}(L_j - 1)\prod_{k\in \text{pa}(j)} L_k$ for cBN.

There also exist copula-based methods (Cui et al., 2016; Castelletti, 2024), which are quite flexible in incorporating different types of data including ordinal data. While the estimation procedures are all very different from each other, the main theoretical difference between the proposed oBN model and the copula-based methods is three-fold:

1. The learned causality or conditional independence is on the observed variable level for the proposed oBN model and is on the latent variable level for the copula-based methods. For

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Table 2: OCD-Depression data. A list of significant edges identified by oBN ranked by inclusion probabilities.

Significant Edge			Probability
obdistress	→	obtime	0.98
middle	→	onset	0.972
compinterf	→	obinterfer	0.946
sad	→	suicide	0.914
concen	→	retard	0.898
middle	→	late	0.862
obinterfer	→	obdistress	0.846
compinterf	→	comptime	0.79
fatigue	→	hypersom	0.758
comptime	→	obtime	0.756
sad	→	anhedonia	0.748
sad	→	guilt	0.722
fatigue	→	retard	0.716
obcontrol	→	obresist	0.67
compresis	→	obresist	0.636
compinterf	→	compcont	0.624
appetite	→	weight	0.596
late	→	hypersom	0.588
anhedonia	→	fatigue	0.58
compinterf	→	compdis	0.554
obdistress	→	obcontrol	0.53
obcontrol	→	compcont	0.53
fatigue	→	concen	0.528
compdis	→	comptime	0.524
obdistress	→	sad	0.514
concen	→	agitation	0.508
compcont	→	compresis	0.504

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

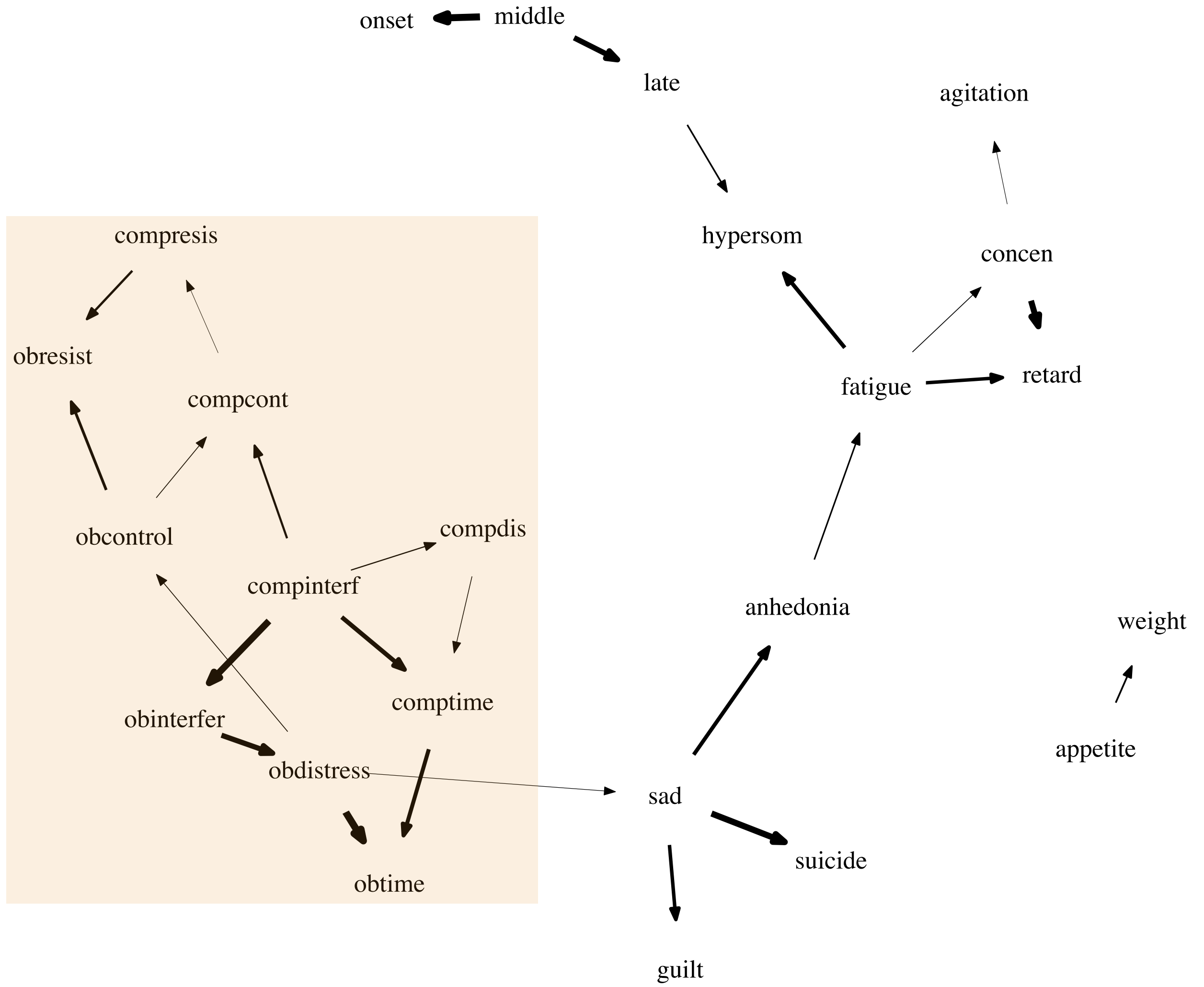


Figure 7: Estimated OCD-Depression networks using oBN with 500 bootstrap samples. The edge width is proportional to its probability. Nodes within the box are the ten OCD-related variables.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

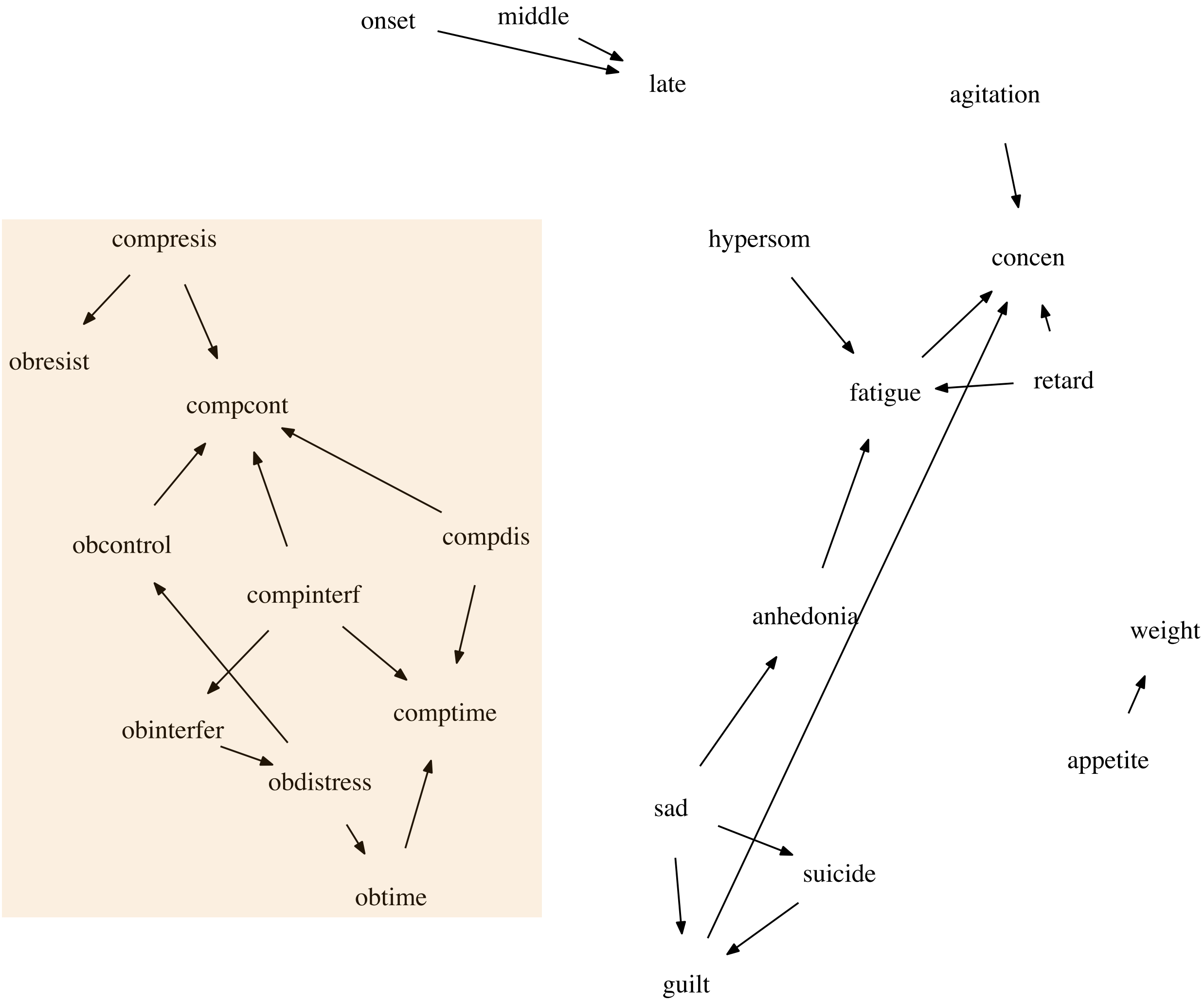


Figure 8: Estimated OCD-Depression networks using PC+oBN. Nodes within the box are the ten OCD-related variables.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

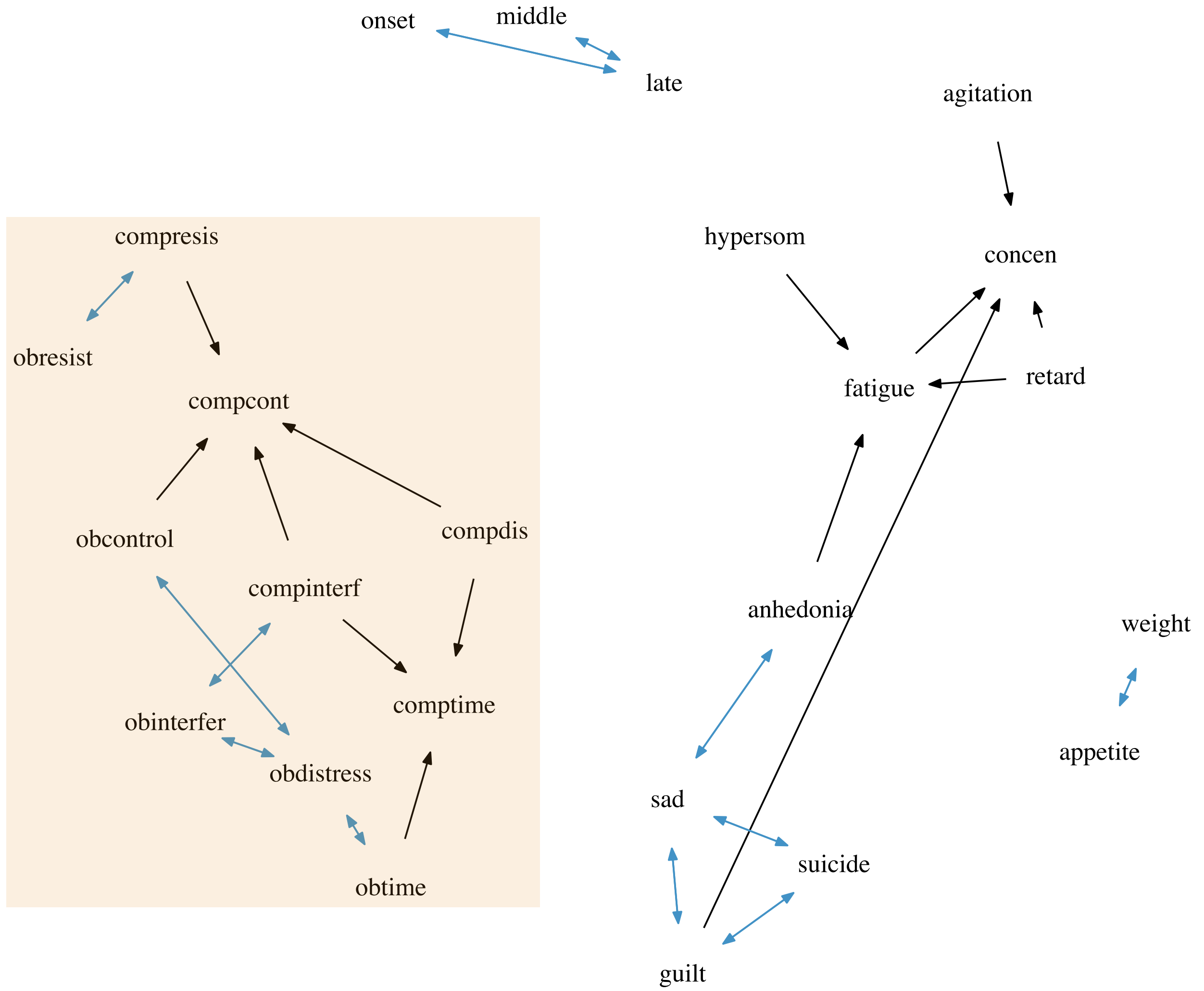


Figure 9: Estimated OCD-Depression networks using PC. The (blue) bidirected edges are edges of which the directionality is undetermined. Nodes within the box are the ten OCD-related variables.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

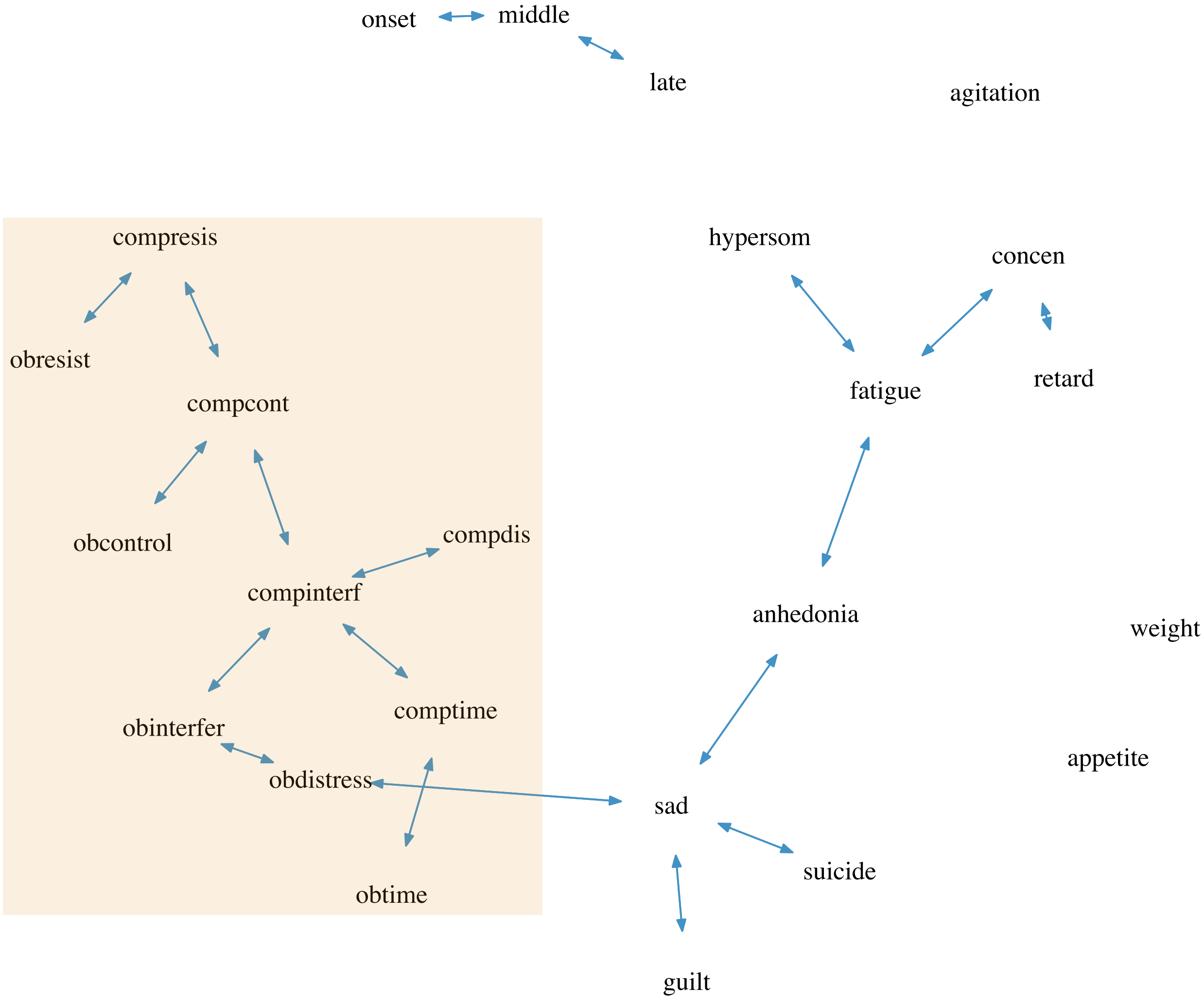


Figure 10: Estimated OCD-Depression networks using cBN with BIC and hill-climbing. The (blue) bidirected edges are edges of which the directionality is undetermined. Nodes within the box are the ten OCD-related variables.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

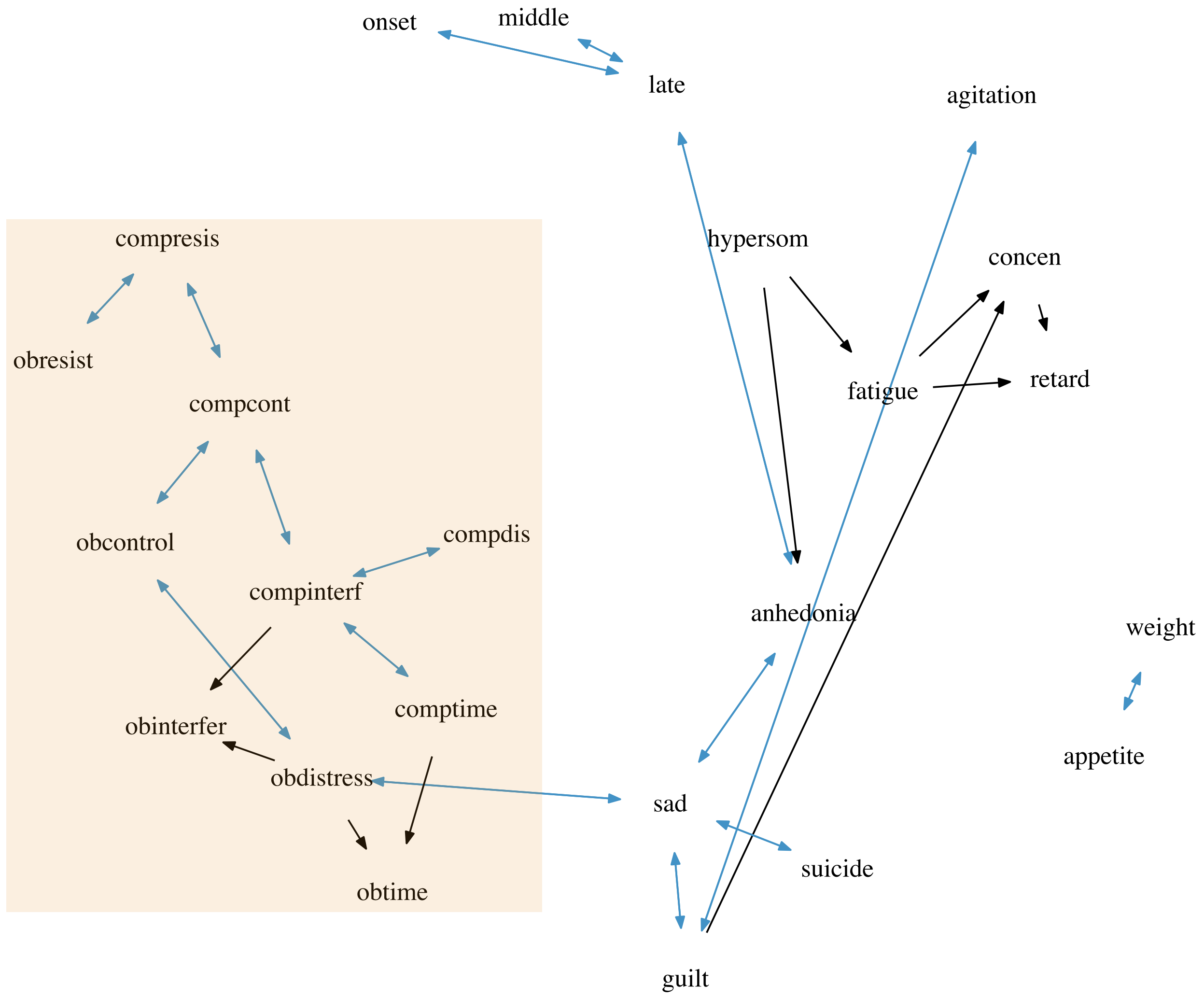


Figure 11: Estimated OCD-Depression networks using OSEM. The (blue) bidirected edges are edges of which the directionality is undetermined. Nodes within the box are the ten OCD-related variables.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

discrete variables, the latter is not equivalent to the former.

2. For mixed data, the proposed oBN model needs to discretize the count/continuous data whereas copula-based methods do not need to because of the latent continuous variable representation.

3. The proposed oBN is uniquely identifiable whereas the copula-based methods are only identifiable up to the Markov equivalence class.

The proposed causal structure learning algorithms based on hill-climbing and bootstrapping worked quite well in simulation studies and also generated some plausible causal hypotheses in the real data. Although this paper focuses on causal structure learning, the discovered DAG structure can be used to determine the causal effects. Following Castelletti et al. (2023), we define the causal effect of X_k on X_j at level $X_j = y$ and $X_k = x$ using $X_k = 1$ as the reference level for $j, k \in V$, $j \neq k$, $y \in \{1, \ldots, L_j\}$, and $x \in \{2, \ldots, L_k\}$ as,

$$\begin{aligned} c_{j,k}(y,x) &= P(X_j = y|\text{do}(X_k = x)) - P(X_j = y|\text{do}(X_k = 1)) \\ &= \sum_{\boldsymbol{z} \in \mathcal{X}_k} P(X_j = y|X_k = x, \boldsymbol{X}_{\text{pa}(k)} = \boldsymbol{z}) P(\boldsymbol{X}_{\text{pa}(k)} = \boldsymbol{z}) \\ &- \sum_{\boldsymbol{z} \in \mathcal{X}_k} P(X_j = y|X_k = 1, \boldsymbol{X}_{\text{pa}(k)} = \boldsymbol{z}) P(\boldsymbol{X}_{\text{pa}(k)} = \boldsymbol{z}), \end{aligned}$$

where $\mathcal{X}_k = \times_{h \in \text{pa}(k)} \{1, \ldots, L_h\}$. Given the estimated DAG and model parameters, the conditional/marginal probabilities needed to compute the causal effect can be calculated using the sum-product message passing algorithm (Koller and Friedman, 2009, Chapter 10).

We hope we have convinced researchers to start using oBNs instead of cBNs or PC for ordinal questionnaire data. There are, of course, limitations of oBNs. First, feedbacks are not allowed in BNs. This may partially explain the inconsistency of some of the causal directions (e.g., the link between *fatigue* and *hypersom*) across the methods in the OCD-Depression data analyses. To infer feedbacks, directed cyclic graphical models may be used. However, we are not aware of the existence of such model for categorical data. Second, the learning algorithm has no guarantee for global convergence. Although Algorithm 2 with multiple random initializations can help mitigate this issue, a more principled solution would be via Bayesian causal structure learning algorithms

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

(Choi et al., 2020), which have theoretical convergence guarantees. We leave it as future work since the proposed algorithms in this paper already showed empirically favorable performance compared to alternative methods. Third, the identifiability theory of Ni and Mallick (2022) requires the causal sufficiency assumption, i.e., there is no unmeasured confounder, of which the validity is difficult to check in practice. Fortunately, their sensitivity analyses provide some assurance that oBN is reasonably robust with respect to the presence of unmeasured confounders. Moreover, the large sample property of oBNs is established under the assumed parametric assumptions. Therefore, users are advised to use domain knowledge to gauge the plausibility of the causal structure. Fourth, oBNs are exploratory, not confirmatory. To confirm the causal hypotheses generated from oBNs, one has to, ultimately, resort to clinical interventions, which is beyond the scope of this paper.

References

Abramowitz, J. S. (2004). Treatment of obsessive-compulsive disorder in patients who have comorbid major depression. *Journal of Clinical Psychology*, 60(11):1133–1141.

Agresti, A. (2003). *Categorical Data Analysis*, volume 482. John Wiley & Sons.

Angst, J., Angst, F., and Stassen, H. H. (1999). Suicide risk in patients with major depressive disorder. *Journal of Clinical Psychiatry*, 60(2):57–62.

Anholt, G. E., Aderka, I. M., Van Balkom, A. J., Smit, J. H., Hermesh, H., De Haan, E., and Van Oppen, P. (2011). The impact of depression on the treatment of obsessive–compulsive disorder: results from a 5-year follow-up. *Journal of Affective Disorders*, 135(1-3):201–207.

Bentler, P. M. and Chou, C.-P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16(1):78–117.

Bird, J. C., Evans, R., Waite, F., Loe, B. S., and Freeman, D. (2019). Adolescent paranoia: prevalence, structure, and causal mechanisms. *Schizophrenia Bulletin*, 45(5):1134–1142.

Brådvik, L. (2018). Suicide risk and mental disorders. *International Journal of Environmental Research and Public Health*, 15(9):2028.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Briganti, G. (2022). On the use of Bayesian artificial intelligence for hypothesis generation in psychiatry. *Psychiatria Danubina*, 34(Suppl 8):201–206.

Briganti, G., Scutari, M., and McNally, R. J. (2022). A tutorial on Bayesian networks for psychopathology researchers. *Psychological Methods*.

Carbonella, J. Y. (2018). *Obsessive-compulsive disorder, trauma, and stress: A network approach*. PhD thesis, University of Miami.

Castelletti, F. (2024). Learning Bayesian networks: a copula approach for mixed-type data. *Psychometrika*, pages 1–29.

Castelletti, F., Consonni, G., and Della Vedova, M. L. (2023). Joint structure learning and causal effect estimation for categorical graphical models. *arXiv preprint arXiv:2306.16068*.

Cervin, M., Lázaro, L., Martínez-González, A. E., Piqueras, J. A., Rodríguez-Jiménez, T., Godoy, A., Aspvall, K., Barcaccia, B., Pozza, A., and Storch, E. A. (2020). Obsessive-compulsive symptoms and their links to depression and anxiety in clinic-and community-based pediatric samples: A network analysis. *Journal of Affective Disorders*, 271:9–18.

Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, 48:55–65.

Choi, J., Chapkin, R., and Ni, Y. (2020). Bayesian causal structural learning with zero-inflated Poisson Bayesian networks. In *Advances in Neural Information Processing Systems 33*.

Choi, J. and Ni, Y. (2023). Model-based causal discovery for zero-inflated count data. *Journal of Machine Learning Research*, 24(200):1–32.

Colombo, D., Maathuis, M. H., et al. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782.

Cui, R., Groot, P., and Heskes, T. (2016). Copula PC algorithm for causal discovery from mixed data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML*

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16, pages 377–392. Springer.

de Wildt, W. A., Lehert, P., Schippers, G. M., Nakovics, H., Mann, K., and van den Brink, W. (2005). Investigating the structure of craving using structural equation modeling in analysis of the obsessive-compulsive drinking scale: A multinational study. *Alcoholism: Clinical and Experimental Research*, 29(4):509–516.

Ebert-Uphoff, I. and Deng, Y. (2012). Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics*, pages 569–593. Springer.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368.

Fried, E. I., van Borkulo, C. D., Cramer, A. O., Boschloo, L., Schoevers, R. A., and Borsboom, D. (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, 52(1):1–10.

Friedman, N., Goldszmidt, M., and Wyner, A. (1999). Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 196–205, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620.

Hong, J. P., Samuels, J., Bienvenu III, O. J., Cannistraro, P., Grados, M., Riddle, M. A., Liang, K.-Y., Cullen, B., Hoehn-Saric, R., and Nestadt, G. (2004). Clinical correlates of recurrent major depression in obsessive–compulsive disorder. *Depression and Anxiety*, 20(2):86–91.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Jöreskog, K. G. (2005). Structural equation modeling with ordinal variables using LISREL. *Technical Report, Scientific Software International, Inc., Lincolnwood, IL.*

Kolenikov, S. (2011). Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology*, 41(1):119–157.

Kolenikov, S. and Bollen, K. A. (2012). Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research*, 41(1):124–167.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* MIT press.

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press.

Lazkano, E., Sierra, B., Astigarraga, A., and Martinez-Otzeta, J. M. (2007). On the use of Bayesian networks to develop behaviours for mobile robots. *Robotics and Autonomous Systems*, 55(3):253–265.

Luo, X. G., Moffa, G., and Kuipers, J. (2021). Learning Bayesian networks from ordinal data. *Journal of Machine Learning Research*, 22(266):1–44.

McNally, R., Mair, P., Mugno, B., and Riemann, B. (2017). Co-morbid obsessive–compulsive disorder and depression: A Bayesian network approach. *Psychological Medicine*, 47(7):1204–1214.

Meyer, J. M., McNamara, J. P., Reid, A. M., Storch, E. A., Geffken, G. R., Mason, D. M., Murphy, T. K., and Bussing, R. (2014). Prospective relationship between obsessive–compulsive and depressive symptoms during multimodal treatment in pediatric obsessive–compulsive disorder. *Child Psychiatry & Human Development*, 45(2):163–172.

Nestadt, G., Samuels, J., Riddle, M., Liang, K.-Y., Bienvenu, O., Hoehn-Saric, R., Grados, M., and Cullen, B. (2001). The relationship between obsessive–compulsive disorder and anxiety and affective disorders: results from the johns hopkins ocd family study. *Psychological Medicine*, 31(3):481–487.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.

Ni, Y. and Mallick, B. (2022). Ordinal causal discovery. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (just-accepted)*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., et al. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5):573–583.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J., and Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, 440(7084):676–679.

Shen, X., Ma, S., Vemuri, P., and Simon, G. (2020). Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific Reports*, 10(1):1–12.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. MIT press.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press

Steketee, G., Frost, R., and Bogart, K. (1996). The Yale-Brown obsessive compulsive scale: Interview versus self-report. *Behaviour Research and Therapy*, 34(8):675–684.

Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, 52(1):313–354.

Verma, T. S. and Pearl, J. (2022). Equivalence and synthesis of causal models. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 221–236.

Vowels, M. J. (2021). Misspecification and unreliable interpretations in psychology and social science. *Psychological Methods*.

Zandberg, L. J., Zang, Y., McLean, C. P., Yeh, R., Simpson, H. B., and Foa, E. B. (2015). Change in obsessive-compulsive symptoms mediates subsequent change in depressive symptoms during exposure and response prevention. *Behaviour Research and Therapy*, 68:76–81.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA. AUAI Press.

Zhou, F., He, K., and Ni, Y. (2023). Individualized causal discovery with latent trajectory embedded bayesian networks. *Biometrics*.

https://doi.org/10.1017/psy.2024.11 Published online by Cambridge University Press