Computational Statistics and Data Analysis 159 (2021) 107209

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Tests for differential Gaussian Bayesian networks based on quadratic inference functions

Xianzheng Huang [a,*], Hongmei Zhang [b]

[a] Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
[b] Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN 38111, USA

ARTICLE INFO

Article history:
Received 12 March 2020
Received in revised form 19 February 2021
Accepted 24 February 2021
Available online 8 March 2021

Keywords:
Directed acyclic graph
Information criterion
Pair bootstrap
Topological ordering
Wild bootstrap

ABSTRACT

Hypotheses testing procedures based on quadratic inference functions are proposed to test whether two Gaussian Bayesian networks are differential in structure, strength of associations between nodes, or both. Bootstrap procedures are developed to estimate *p*-values to quantify the statistical significance of the tests. Operating characteristics of these testing procedures are investigated using synthetic data in simulation experiments. Additionally, the proposed methods are applied to flow cytometry data from a designed experiment, and data of bile acids from an observational study in the Alzheimer's Disease Neuroimaging Initiative.

1. Introduction

There has been an explosion of interest on graphical models among data scientists in the past two decades. Graphs used for visualizing relationships between variables were initially formulated and developed within the artificial intelligence community, and quickly attracted attention of researchers in a wide range of scientific fields (Jensen, 1996; Lauritzen, 1996; Edwards, 2012; Neapolitan, 2012; Pearl, 2014). A graph consists of nodes, i.e., vertices, representing variables of interest in a system, the relationships between which are depicted via edges connecting nodes. Formulating a probabilistic graphical model involves the specification of a graph structure and a probability model for the nodes. A graphical model with undirected edges is an undirected network, also referred to as Markov network, which encodes the conditional dependence relationships between nodes. A graphical model with directed edges is a Bayesian network, also referred to as a directed acyclic graph (DAG) to signify that there exists no path that starts from one node and ends at the same node. When there is an edge pointing from one node to another node, these two nodes are referred to as a parent node (of the latter) and a child node (of the former), respectively. A Bayesian network satisfies the local Markov property in the sense that, given its parents, a node is independent of its non-descendant nodes. Hence, a Bayesian network specifies a factorization of the joint distribution of all nodes in the graph. Following this factorization, one can more efficiently implement probabilistic inference on causal relationships between the nodes, and further deduce conditional dependence relationships.

Applications of Bayesian networks include profiling gene maps in genetics studies, predicting a treatment outcome in the medical field, and conducting financial analysis in econometrics, among many other examples. In particular, identifying differential Bayesian networks is of great interest to many domain scientists. For example, comparing a cellular signaling

* Corresponding author.
E-mail address: huangt@stat.sc.edu (X. Huang).

https://doi.org/10.1016/j.csda.2021.107209
0167-9473/

network associated with a diseased population and its counterpart network associated with a healthy population can provide insight on the impact of the disease on the concert work of relevant cells. In this article, we develop methods to infer whether or not a Gaussian Bayesian network is differential between two populations. The study presented here distinguish from existing relevant works in at least three aspects. First, we test differential directed networks as opposed to differential undirected networks, with a sizable collection of existing literature on the latter yet very limited research on the former. For instance, Gill et al. (2010) proposed a procedure to globally test differential undirected graphs based on strength of genetic associations or interaction between genes. Jacob et al. (2012) tested multivariate two-sample means associated with graphs of known structures using Hotelling's T^2-tests. Zhao et al. (2014) developed a method to infer differences in two precision matrices corresponding to two undirected networks. Xia et al. (2015) extended their work in order to globally test differentiation of undirected graphs. Städler et al. (2017) also developed methods for testing differentiations of undirected graphs based on precision matrices. Durante et al. (2018) studied associations of undirected networks with a feature of interest. Zhao et al. (2019) developed a general framework for testing differential connectivity between two undirected networks, which aims to qualitatively compare structures of two precision matrices instead of quantitatively comparing entries of them. This highlighted feature in Zhao et al. (2019) relates to the second aspect that makes our methods stand out from existing works, which is that our proposed testing procedures can test differentiation in regard to solely graph structure, or strengths of associations between variables, or both. Third, the proposed inference procedures are applicable to data from an observational study or from designed experiments. Drawing inference for a Bayesian network can be more challenging than for an undirected network because a Bayesian network can be identified based on data from an observational study only up to a Markov equivalence class (Verma and Pearl, 1991; Andersson et al., 1997; Chickering, 2002; Hauser and Bühlmann, 2012). Inclusion of interventional data from a designed experiment improves the identifiability of a Bayesian network. In such an experiment, one sets the values of some node(s) to be pre-specified values, which in effect destroys the causal dependencies of the intervened node(s). Ellis and Wong (2008) developed a fast MCMC algorithm based on experimental data that include interventional data and observational data. Besides inclusion of interventional data, incorporating information of the topological ordering of nodes also improves identifiability. Given a directed graph structure, a topological ordering specifies a sequence of nodes such that a child node in the graph always comes after its parent nodes (Cormen et al., 2001, Section 22.4). We do not assume ordering known in our study. Lastly, we formulate and infer Bayesian networks under the regression framework with Gaussian model error, a framework that differs from most frameworks adopted in the artificial intelligence community (e.g., Chung et al., 2006; Nielsen and Jensen, 2009), where one directly models edge probabilities in a Bayesian network for instance.

To prepare for methodology development, we first formulate regression models that characterize a Bayesian network in Section 2. Section 3 presents an algorithm for inferring a network based on penalized score equations. Test statistics based on quadratic inference functions are proposed in Section 4 for testing hypotheses relating to comparisons between two networks. To quantify statistical significance of the proposed test statistics, we develop bootstrap procedures in Section 5 to estimate the null distribution of the test statistics. Operating characteristics of the testing procedures are investigated via simulation experiments reported in Section 6. The proposed methods are applied to two real-life applications in Section 7. Lastly, we summarize the contribution of the current study and discuss follow-up research agenda in Section 8.

2. Model and data

Suppose that the same set of p nodes are considered for two populations. It is of interest to compare the underlying Bayesian networks involving this set of nodes associated with the two populations. Refer to node j of population k as $X_j^{(k)}$, for $j = 1, \ldots, p$, $k = 1, 2$. Suppose that associations between nodes in population k is specified by p linear regression models,

$$X_j^{(k)} = \sum_{j'=1}^{p} X_{j'}^{(k)} \mathbf{B}^{(k)}[j', j] + \epsilon_j^{(k)}, \text{ for } j = 1, \ldots, p, \tag{1}$$

where $\epsilon_j^{(k)} \sim N(0, \sigma_{kj}^2)$ is the model error independent of $X_{j'}^{(k)}$ for $j' \neq j$, $\mathbf{B}^{(k)}$ is the $p \times p$ matrix of regression coefficients, of which the diagonal entries are equal to zero. We assume that each node has mean zero so that no intercept is needed in (1). Note that $X_{j'}^{(k)}$ is a parent of $X_j^{(k)}$ if and only if $\mathbf{B}^{(k)}[j', j] \neq 0$, thus $\mathbf{B}^{(k)}$ fully determines the graph structure, denoted by G_k. Having the jth column $\mathbf{B}^{(k)}[\,, j]$ equal to $\mathbf{0}_{p\times 1}$ indicates that $X_j^{(k)}$ is parentless, in which case we call it a root node. Having the jth row $\mathbf{B}^{(k)}[j, \,]$ equal to $\mathbf{0}_{p\times 1}^{\mathrm{T}}$ means that $X_j^{(k)}$ is childless. Define $\mathbf{X}^{(k)}$ as an $N_k \times p$ matrix that stores a data set of size N_k from population k, for $k = 1, 2$.

The data may come from an observational study or a designed experiment. In the latter case, certain nodes are intentionally inhibited or stimulated under different experimental conditions. To allow data from either type of study, we let $\mathscr{I}_{kj}$, with cardinality n_{kj}, be the index set relating to rows of $\mathbf{X}^{(k)}$ that contain interventional data for $X_j^{(k)}$, and let $\mathscr{O}_{kj}$, with cardinality $n_{-kj} = N_k - n_{kj}$, be the index set corresponding to the remaining rows in $\mathbf{X}^{(k)}$ where observational data of $X_j^{(k)}$ are stored. If data are from an observational study, or they are from a designed experiment where node j is never intervened, then $\mathscr{I}_{kj}$ is the empty set and $\mathscr{O}_{kj} = \{1, \ldots, N_k\}$. When considering the jth regression model with

$X_j^{(k)}$ as the response variable, one should only use the observational data associated with $X_j^{(k)}$, i.e., $\mathbf{X}^{(k)}[\mathscr{O}_{kj},]$, because any potential causal effect of other nodes on $X_j^{(k)}$ is suppressed by design in $\mathbf{X}^{(k)}[\mathscr{I}_{kj},]$.

In this study, we develop procedures for testing hypotheses in regard to how the two networks compare based on $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Hypotheses of research interest include $H_0^a : G_1 = G_2$ versus $H_1^a : G_1 \neq G_2$ when the comparison focuses solely on the graph structure, and a more detailed comparison regarding the strength of associations between nodes with the null hypothesis $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, and the alternative hypothesis $H_1^b : \mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$. Clearly, H_1^a being true implies that H_0^b is false, but not vice versa. To reject H_0^b but not H_0^a is to conclude that there exist associations between nodes that are differential between two networks sharing a common graph structure. When testing these hypotheses, graph estimation is involved, for which we describe an algorithm next.

3. Graph estimation

In this section, we suppress the population index k from the notations defined in Section 2 and focus on estimating one graph using one data matrix in general. For instance, $\mathbf{X}$ is an $N \times p$ data matrix available for inferring causal relationships between p nodes, and $\mathbf{B}$ is the $p \times p$ matrix of regression coefficients corresponding to the network. For node X_j, there are $n_j (\geq 0)$ rows composed of interventional data with row index set $\mathscr{I}_j$; the rest of $n_{-j} = N - n_j$ rows, corresponding to the index set $\mathscr{O}_j$, contain observational data for X_j, for $j = 1, \ldots, p$.

3.1. Parent selection

Under the regression framework, inferring a Bayesian network with p nodes translates to inferring p regression models in (1) simultaneously under the acyclic constraint. To facilitate selecting parents for each node, we apply the score-based variable selection method proposed by Huang and Zhang (2013) in conjunction with cycle elimination via Kahn's algorithm (Kahn, 1962). The following algorithm describes this estimation procedure, with Kahn's algorithm relegated to Section 3.2.

Step 1: Compute the least squares estimate of $\mathbf{B}_j = \mathbf{B}[-j, j]$ based on data $\mathbf{X}[\mathscr{O}_j,]$. Denote by $\hat{\mathbf{B}}_j^{(0)}$ this estimate, for $j = 1, \ldots, p$. Set the iteration index $t = 0$.

Step 2: For $j = 1, \ldots, p$, use $\hat{\mathbf{B}}_j^{(t)}$ as the starting value to solve the following penalized score estimating equations,

$$n_{-j}^{-1} \sum_{\ell \in \mathscr{O}_j} \Psi_{j\ell}(\mathbf{B}_j) - \tilde{P}_\lambda(\mathbf{B}_j) = \mathbf{0}, \tag{2}$$

where

$$\Psi_{j\ell}(\mathbf{B}_j) = (\mathbf{X}[\ell, j] - \mathbf{X}[\ell, -j]\mathbf{B}_j)\mathbf{X}[\ell, -j]^{\mathrm{T}}, \text{ for } \ell \in \mathscr{O}_j, \tag{3}$$

is the normal score associated with the jth regression model evaluated at one data point, and $\tilde{P}_\lambda(\mathbf{B}_j)$ is the partial derivative of the SCAD penalty function (Fan and Li, 2001),

$$P_\lambda(t) = \lambda t I(0 \leq t < \lambda) + \frac{(a^2-1)\lambda^2 - (t - a\lambda)^2}{2(a-1)} I(\lambda \leq t < a\lambda) + \frac{(a+1)\lambda^2}{2} I(t \geq a\lambda),$$

in which λ is a tuning parameter, $a = 3.7$, and $I(\cdot)$ is the indicator function. More specifically, entries of the $(p-1)\times 1$ vector $\tilde{P}_\lambda(\mathbf{B}_j)$ are given by, for $i \neq j$,

$$\frac{\partial}{\partial \beta_{ij}} P_\lambda(|\beta_{ij}|) = \lambda \left\{ I(|\beta_{ij}| \leq \lambda) + \frac{(a\lambda - |\beta_{ij}|)_+}{(a-1)\lambda} \right\} \operatorname{sign}(\beta_{ij}),$$

where $\beta_{ij} = \mathbf{B}[i, j]$. The p sets of solutions to (2) give the p columns of an updated matrix of regression coefficients estimates, $\hat{\mathbf{B}}^{(t+1)}$. Denote by $\tilde{G}$ the graph structure indicated by $\hat{\mathbf{B}}^{(t+1)}$.

Step 3: For $i, j = 1, \ldots, p$ and $i \neq j$, compute the unpenalized estimate of β_{ij}, which is the least squares estimate, if X_i is a parent of X_j according to $\tilde{G}$, and obtain the estimated standard error of the unpenalized estimate via sandwich variance estimation for M-estimators. Compute the p-value based on the least squares estimate of β_{ij} along with its estimated standard error for testing $H_0 : \beta_{ij} = 0$ versus $H_1 : \beta_{ij} \neq 0$.

Step 4: Implement Kahn's algorithm to inspect and eliminate cycles in $\tilde{G}$ based on the p-values from Step 3. More entries in $\hat{\mathbf{B}}^{(t+1)}$ are set to zero during this cycle elimination process unless no cycle is detected. More details of this step are given in Section 3.2.

Step 5: If $|\hat{\mathbf{B}}^{(t+1)} - \hat{\mathbf{B}}^{(t)}|_\infty > 10^{-4}$, set $t = t + 1$, and return to Step 2. Otherwise, output $\hat{\mathbf{B}}^{(t+1)}$ as the final estimate of $\mathbf{B}$, denoted by $\hat{\mathbf{B}}$, and denote the corresponding graph structure as $\hat{G}$.

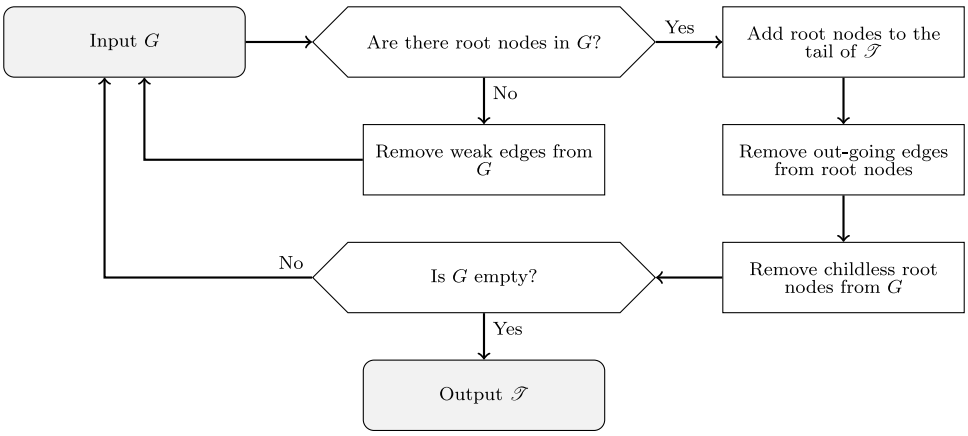


Fig. 1. The algorithm for eliminating cycles in G and finding a topological ordering, $\mathscr{T}$, compatible with the resultant acyclic G.

In Step 5, for a matrix $\mathbf{A}$, $|\mathbf{A}|_\infty$ denotes the largest entry of $\mathbf{A}$ in absolute value. In Step 2, one may consider other penalty functions such as the LASSO (Tibshirani, 1996) and the adaptive LASSO (ALASSO) (Zou, 2006) in place of SCAD when constructing the penalized estimating equations in (2). Both SCAD and ALASSO have been shown to enjoy the appealing oracle properties in variable selections. Between the two penalty functions, we choose SCAD in order to avoid specifying the adaptive weights required in ALASSO, and also because Aragam and Zhou (2015) showed that a concave penalty, such as SCAD, offers improved performance in Bayesian network structure learning when comparing with an L_1-based penalty. To choose a value for the tuning parameter λ in (2), we adopt the score-based information criterion inspired by the following quadratic inference function (Lindsay and Qu, 2003),

$$Q_j = \left\{n_{-j}^{-1}\Psi_j(\mathbf{B}_j)\right\}^{\mathrm{T}} \left\{\mathbf{H}_j(\mathbf{B}_j)\right\}^{-1} \left\{n_{-j}^{-1}\Psi_j(\mathbf{B}_j)\right\}, \tag{4}$$

where $\Psi_j(\mathbf{B}_j) = \sum_{\ell\in\mathscr{O}_j} \Psi_{j\ell}(\mathbf{B}_j)$ and $\mathbf{H}_j(\mathbf{B}_j) = n_{-j}^{-1}\sum_{\ell\in\mathscr{O}_j} \Psi_{j\ell}(\mathbf{B}_j)\Psi_{j\ell}(\mathbf{B}_j)^{\mathrm{T}}$. In particular, we choose λ so that the following score-based information criterion evaluated $\hat{\mathbf{B}}$ is minimized,

$$\mathrm{SIC}(\hat{\mathbf{B}}) = \sum_{j=1}^{p}\left(\hat{Q}_j + e_j\frac{\log n_{-j}}{n_{-j}}\right), \tag{5}$$

where the dependence of $\hat{\mathbf{B}}$ on λ is suppressed but implied by the above algorithm, e_j is the number of parents of X_j according to $\hat{\mathbf{B}}$, and $\hat{Q}_j$ is equal to Q_j in (4) evaluated at the unpenalized estimate of $\mathbf{B}_j$, denoted by $\tilde{\mathbf{B}}_j$, which is the least squares estimate of $\mathbf{B}_j$ given the graph structure $\hat{G}$. It is shown in Huang and Zhang (2021) that $\mathrm{SIC}(\cdot)$ is a consistent information criterion under mild regularity conditions.

3.2. Cycle elimination

In Step 4 of the graph estimation algorithm in Section 3.1, we use a topological sorting algorithm, known as Kahn's algorithm, to detect and delete cycles in $\tilde{G}$. Given a generic directed graph, there may exist more than one topological ordering compatible with it. For example, if there are two adjacent nodes in a sorted sequence that are not connected by an edge in the graph, then swapping these two nodes yields a different ordering that is compatible with the same graph. A sorting algorithm typically is an iterative algorithm with a built-in check for cycles in a directed graph. Fig. 1 illustrates the sorting algorithm we use to revise an initial directed graph structure G to yield an acyclic graph, during the process of which a topological ordering of the nodes compatible with the final acyclic G is found. Before starting the algorithm, one sets up a queue, denoted by $\mathscr{T}$, initially empty, for storing the sorted nodes as they come in.

Applying the algorithm depicted in Fig. 1 to $\tilde{G}$ in Step 4 of the graph estimation algorithm, we use $\hat{\mathbf{B}}^{(t+1)}$ resulting from Step 2 to monitor updates in $\tilde{G}$ in four ways. First, to find a root node, one simply looks for a column in $\hat{\mathbf{B}}^{(t+1)}$ that contains all zeros. Suppose $\hat{\mathbf{B}}^{(t+1)}[\,,j] = \mathbf{0}_{p\times 1}$, then X_j is a root node. Second, to remove out-going edges from a root node amounts to setting all entries in the row corresponding to this node at zero, which effectively makes this node childless. Suppose X_j is a root node, then one sets $\hat{\mathbf{B}}^{(t+1)}[j,\,] = \mathbf{0}_{p\times 1}^{\mathrm{T}}$ to make X_j childless. Third, removing a childless root node from $\tilde{G}$ is equivalent to deleting the row and the column corresponding to this node from $\mathbf{B}$. Lastly, a weak edge in $\tilde{G}$ corresponds to a non-zero entry in the current $\hat{\mathbf{B}}^{(t+1)}$ whose p-value is the largest among the p-values associated with remaining nodes obtained in Step 3, since a larger p-value for β_{ij} implies a weaker association between X_i and X_j. Removing a weak edge is to set this entry in $\hat{\mathbf{B}}^{(t+1)}$ at zero.

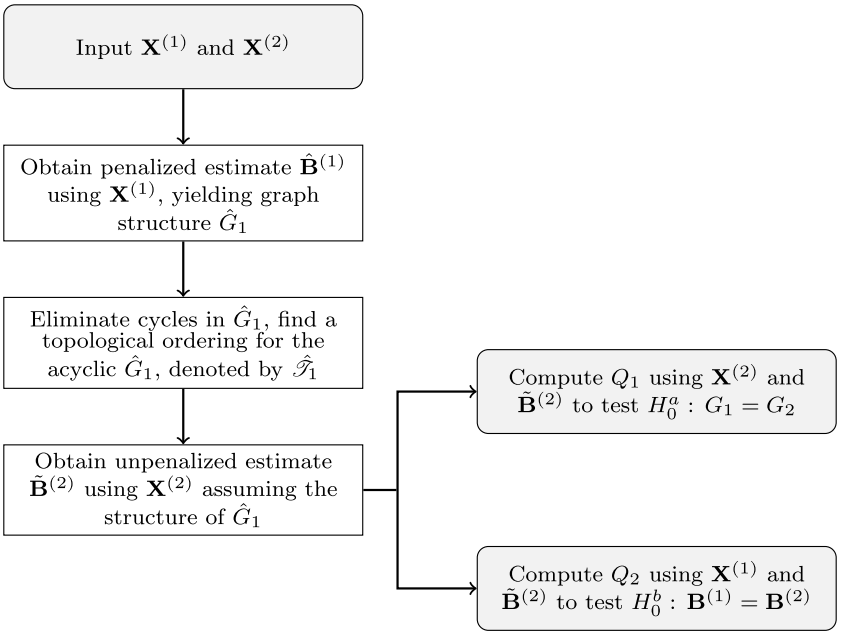


Fig. 2. Construction of the two test statistics, Q_1 for testing H_0^a and Q_2 for testing H_0^b, based on data $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

By the time the queue $\mathscr{T}$ collects all p nodes, $\tilde{G}$ becomes a graph without a node, i.e., an empty graph. It is at this point when we retrieve an acyclic $\tilde{G}$ by deleting from the initial $\tilde{G}$ (resulting from Step 2) the weak edges that are removed during the sorting process depicted in Fig. 1. In other words, the outgoing edges from a root nodes being removed to create childless root nodes, and the childless root nodes being removed during the iterative algorithm are put back to recover the final acyclic $\tilde{G}$, and so are the corresponding rows and columns of $\hat{\mathbf{B}}^{(t+1)}$. The computation time for this algorithm is of order $O(p + |\mathscr{E}_{\tilde{G}}\tilde{G}|)$, where $|\mathscr{E}_{\tilde{G}}\tilde{G}|$ denotes the number of edges in the initial $\tilde{G}$ resulting from Step 2.

4. Test statistics

Equipped with a way to estimate a Bayesian network, we are now ready to construct test statistics for testing $H_0^a : G_1 = G_2$ and $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$. The quadratic inference function in (4) is the building block of these test statistics. For ease of exposition and comparison, Fig. 2 provides a flowchart for the construction of two test statistics, Q_1 and Q_2, to be defined in the upcoming subsections.

4.1. Testing H_0^a

To test $H_0^a : G_1 = G_2$, we define the following test statistic,

$$Q_1 = \sum_{j=1}^{p} \left\{ n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \boldsymbol{\Psi}_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \right\}^{\mathrm{T}} \left\{ \mathbf{H}_j(\tilde{\mathbf{B}}_j^{(2)}) \right\}^{-1} \left\{ n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \boldsymbol{\Psi}_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \right\}, \tag{6}$$

where $\mathbf{H}_j(\tilde{\mathbf{B}}_j^{(2)}) = n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \boldsymbol{\Psi}_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \boldsymbol{\Psi}_{j\ell}(\tilde{\mathbf{B}}_j^{(2)})^{\mathrm{T}}$, $\boldsymbol{\Psi}_{j\ell}(\tilde{\mathbf{B}}_j^{(2)})$ is the normal score evaluated at $\tilde{\mathbf{B}}_j^{(2)}$ and $\mathbf{X}^{(2)}[\ell, \,]$, and $\tilde{\mathbf{B}}_j^{(2)} = \tilde{\mathbf{B}}^{(2)}[-j, j]$ is the unpenalized estimate of the regression coefficients for the jth regression model computed using $\mathbf{X}^{(2)}[\mathscr{O}_{2j}, \,]$ while assuming the graph structure $\hat{G}_1$ obtained from applying the graph estimation algorithm to $\mathbf{X}^{(1)}$.

Under $H_0^a : G_1 = G_2$, $\hat{G}_1$ is also a sensible estimate for G_2, despite how $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ compare quantitatively. Hence, under H_0^a, assuming $\hat{G}_1$ while estimating $\mathbf{B}^{(2)}$ based on $\mathbf{X}^{(2)}$ amounts to estimating $\mathbf{B}^{(2)}$ using data from the second population while assuming an asymptotically correct model (i.e., graph structure) for this population. Here, asymptotics relate to settings with p fixed and the amount of data information available for inferring each regression model diverge. By Lindsay and Qu (2003), as $\min_{j=1,\ldots,p} n_{-2j} \to \infty$, Q_1 is asymptotically the sum of p dependent χ^2 random variables, whose realizations tend to be smaller compared to when a wrong model is assumed for the second population when estimating $\mathbf{B}^{(2)}$. In conclusion, Q_1 is a sensible statistic for testing $H_0^a : G_1 = G_2$ versus $H_1^a : G_1 \neq G_2$, with a larger value of Q_1 providing more evidence in favor of H_1^a, regardless of the quantitative comparison between $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$.

4.2. Testing H_0^b

By revising the test statistic Q_1 in (6), we obtain a test statistic for testing $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$ as follows,

$$Q_2 = \sum_{j=1}^{p} \left\{ n_{-1j}^{-1} \sum_{\ell \in \mathscr{O}_{1j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \right\}^{\mathrm{T}} \left\{ \mathbf{H}_j^*(\tilde{\mathbf{B}}_j^{(2)}) \right\}^{-1} \left\{ n_{-1j}^{-1} \sum_{\ell \in \mathscr{O}_{1j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \right\}, \tag{7}$$

where $\mathbf{H}_j^*(\tilde{\mathbf{B}}_j^{(2)}) = n_{-1j}^{-1} \sum_{\ell \in \mathscr{O}_{1j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(2)}) \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(2)})^{\mathrm{T}}$, in which $\Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(2)})$ is the normal score evaluated at $\tilde{\mathbf{B}}_j^{(2)}$ and $\mathbf{X}^{(1)}[\ell, \ ]$, and $\tilde{\mathbf{B}}_j^{(2)}$ is the same as that appearing in Q_1.

If $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$ is true, $\tilde{\mathbf{B}}^{(2)}$ is also a consistent estimator for $\mathbf{B}^{(1)}$. Therefore, Q_2 is a quadratic inference function evaluated at data from the first population, $\mathbf{X}^{(1)}$, and a consistent estimator of $\mathbf{B}^{(1)}$. Consequently, Q_2 is asymptotically the sum of p dependent χ^2 random variables (Lindsay and Qu, 2003) under H_0^b. If H_0^b is not true, evaluated at $\mathbf{X}^{(1)}$ and an inconsistent estimate of $\mathbf{B}^{(1)}$, i.e., $\tilde{\mathbf{B}}^{(2)}$, Q_2 is asymptotically the sum of p dependent non-central χ^2 random variables that can be much larger than what is expected under H_0^b.

In conclusion, Q_2 can distinguish between $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$ and $H_1^b : \mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$, with a larger value of Q_2 suggesting more evidence supporting H_1^b, despite how the two graph structures compare. In Fig. 2, the topological ordering produced during the graph estimation procedures, $\hat{\mathscr{T}}_1$, will be needed for estimating the null distributions of these test statistics, as elaborated in the next section.

5. Bootstrap procedures

The distributions of Q_1 and Q_2 defined in Section 4 under H_0^a or H_0^b are intractable due to the complicated correlation between the p quadratic forms as the summands of these test statistics. In this section, we develop bootstrap procedures, following the idea of wild bootstrap (Wu, 1986) and pairs bootstrap (Freedman et al., 1981), to estimate the null distributions of the proposed test statistics.

5.1. Wild bootstrap for Q_1 and Q_2

As a parametric bootstrap strategy, wild bootstrap is widely applicable in regression settings. All existing variants of wild bootstrap procedures involve repeatedly generating residuals under certain parametric model assumptions. In the context of graph estimation, where a Gaussian Bayesian network is decomposed into p correlated regression models in (1), we develop wild bootstrap procedures to generate bootstrap analogues of the proposed test statistics under a null hypothesis. Because each test statistic is the sum of correlated quadratic functions, we use a gamma distribution to approximate the null distribution of a proposed test statistic, which is supported by empirical evidence presented in Section 6 (see Figs. 4 and 5).

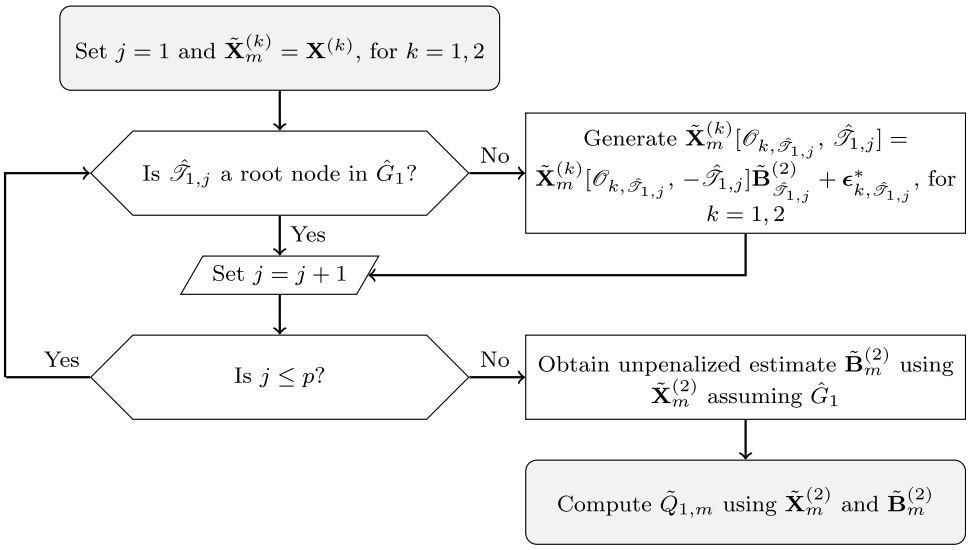
Fig. 3-(a) demonstrates such a wild bootstrap procedure for generating one copy of Q_1 in the bootstrap world, denoted by $\tilde{Q}_{1,m}$. This procedure is repeated M times, yielding $\{\tilde{Q}_{1,m}\}_{m=1}^M$. We then estimate the shape and scale parameters of a gamma distribution via the method of moments using the sample mean and sample variance of $\{\tilde{Q}_{1,m}\}_{m=1}^M$. An estimated p-value associated with Q_1 is defined as the probability that a random variable following this estimated gamma distribution exceeds Q_1.

A few remarks about this wild bootstrap procedure are in order. First, when raw data are from a designed experiment, the bootstrap data $\tilde{\mathbf{X}}_m^{(2)}$ keep the original interventional data from $\mathbf{X}^{(2)}$. This is in the same spirit as using the same design matrix when generating bootstrap response data in the regression setting according to wild bootstrap. Second, observational data associated with different nodes are generated following the ordering of these nodes specified by $\hat{\mathscr{T}}_1$. This shares the same rationale as generating bootstrap time series data recursively following the ordering in time. Third, $\tilde{\mathbf{X}}_m^{(2)}$ keeps the same data from $\mathbf{X}^{(2)}$ for root nodes. This is because root node data are equivalent to data at the initial time point in time series, which are kept unchanged and used to start the recursion for generating bootstrap time series data. Finally, when generating the observational data in $\tilde{\mathbf{X}}_m^{(2)}$ associated with node $\hat{\mathscr{T}}_{1,j}$, the model errors, $\boldsymbol{\epsilon}^*_{2,\hat{\mathscr{T}}_{1,j}}$, are generated from $N(0, \hat{\sigma}^2_{2,\hat{\mathscr{T}}_{1,j}})$, where $\hat{\sigma}^2_{2,\hat{\mathscr{T}}_{1,j}}$ is the sample variance of $n_{-2,\hat{\mathscr{T}}_{1,j}}$ residuals in $\{\mathbf{X}^{(2)}[\ell, \hat{\mathscr{T}}_{1,j}] - \mathbf{X}^{(2)}[\ell, -\hat{\mathscr{T}}_{1,j}]\tilde{\mathbf{B}}^{(2)}_{\hat{\mathscr{T}}_{1,j}\hat{\mathscr{T}}_{1,j}}\}_{\ell \in \mathscr{O}_{2,\hat{\mathscr{T}}_{1,j}}}$. Here, we use $\hat{\mathscr{T}}_{1,j}$ to refer to the jth node in the sorted sequence of nodes according to $\hat{\mathscr{T}}_1$, for $j = 1, \ldots, p$.

Because the observational data in $\tilde{\mathbf{X}}_m^{(2)}$ are regenerated using $\tilde{\mathbf{B}}^{(2)}$, along with the root node data and the interventional data (when they exist) in $\mathbf{X}^{(2)}$, $\tilde{\mathbf{X}}_m^{(2)}$ can be viewed as a bootstrap analogue of $\mathbf{X}^{(2)}$ provided that $\hat{G}_1$ is a consistent estimate of G_2, which is the case under $H_0^a : G_1 = G_2$. Consequently, under H_0^a, $\tilde{Q}_{1,m}$ is a bootstrap analogue of Q_1 even if $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$.

Fig. 3-(b) depicts a similar wild bootstrap procedure for generating a bootstrap copy of Q_2, denoted by $\tilde{Q}_{2,m}$, for $m = 1, \ldots, M$. An estimated p-value associated with Q_2 is similarly obtained as described for Q_1.

(a) Wild bootstrap for Q_1



(b) Wild bootstrap for Q_2

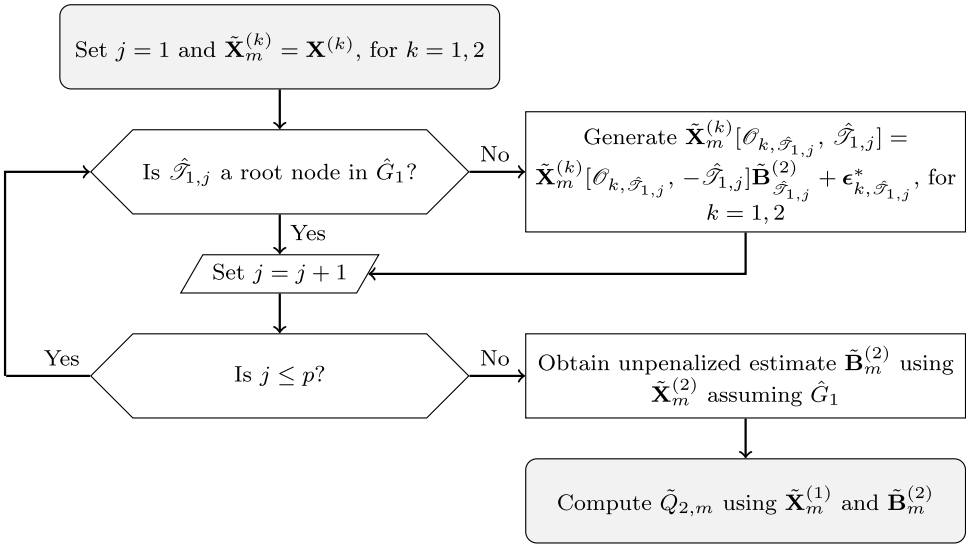


Fig. 3. Wild bootstrap procedures that generate an analogue of Q_1 under $H_0^a : G_1 = G_2$, and an analogue of Q_2 under $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$. Given the topological ordering $\hat{\mathscr{T}}_1$, $\hat{\mathscr{T}}_{1,j}$ refers to the jth node in the ordering.

5.2. Pairs bootstrap for Q_2

Pairs bootstrap is a nonparametric resampling strategy that does not require regenerating data based on certain parametric model assumptions. The data generating scheme is to sample response data and covariates data together in pairs. Besides simplicity, one motivation of this scheme is to preserve the correlation between the response and covariates in the bootstrap data. A criticism of this scheme is that the designed matrix (based on the sampled covariate data) differs from bootstrap sample to bootstrap sample, which is not appealing when one aims to infer the distribution of a response conditioning on the original design matrix. Because of this criticism, some believe that pairs bootstrap is less suitable for inferring the conditional distribution of a response given covariates, and more adequate for drawing inference for the joint distribution of them. We are less concerned about this criticism because, for the purpose of comparing two networks, we need to infer the joint distribution of p nodes, even though we decompose a network into p regression models, each of which specifies a conditional distribution of one node given the rest.

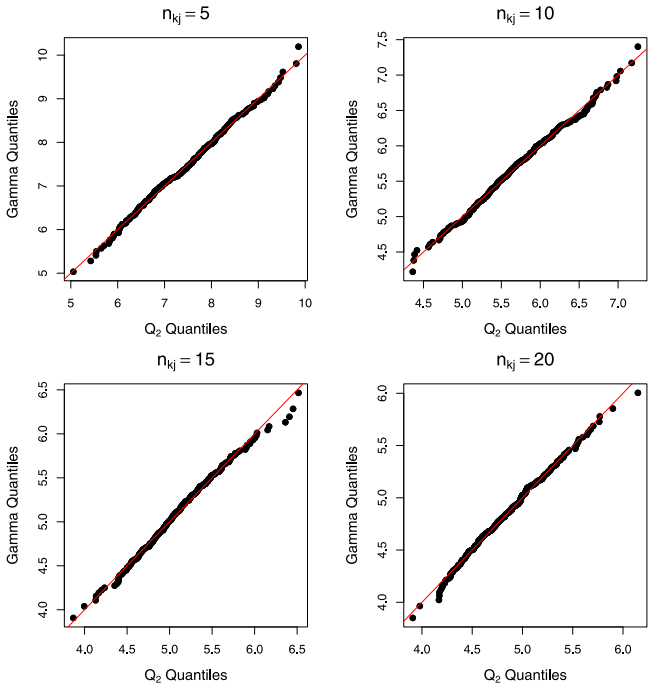


Fig. 4. QQ plots of Q_2 from 300 Monte Carlo replicates generated according to a given graph configuration under (I): $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$.

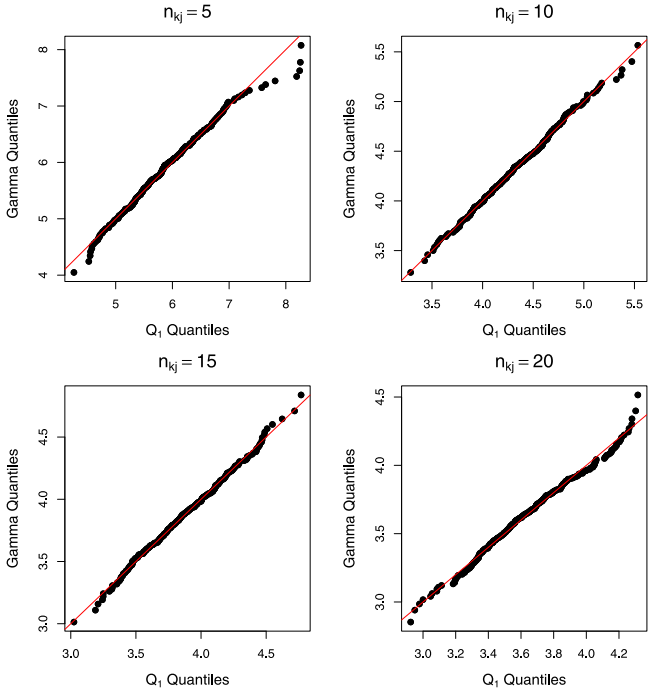


Fig. 5. QQ plots of Q_1 from 300 Monte Carlo replicates generated according to a given graph configuration under (III): $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ and $G_1 = G_2$.

Noting that, under $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, the data generating process leading to $\mathbf{X}^{(1)}$ is the same as that producing $\mathbf{X}^{(2)}$, we propose a pairs bootstrap procedure to approximate the distribution of Q_2 under H_0^b that involves sampling from the $(N_1 + N_2) \times p$ combined data matrix defined as $\mathbf{X}_c = \mathbf{X}^{(1)} \cup \mathbf{X}^{(2)}$. If data are from observational studies, we simply

randomly permute $N_1 + N_2$ rows of $\mathbf{X}_c$ in each round of bootstrap indexed by $m \in \{1, \ldots, M\}$, and then use the first N_1 rows of the resultant combined data matrix to be a bootstrap version of $\mathbf{X}^{(1)}$, denoted by $\tilde{\mathbf{X}}_m^{(1)}$, and use the remaining N_2 rows to be $\tilde{\mathbf{X}}_m^{(2)}$, a bootstrap version of $\mathbf{X}^{(2)}$. If data are from designed experiments, we first permute $n_{1j} + n_{2j}$ rows within $\mathbf{X}_c[\mathscr{I}_{c,j},] = \mathbf{X}^{(1)}[\mathscr{I}_{1j},] \cup \mathbf{X}^{(2)}[\mathscr{I}_{2j},]$, for $j = 1, \ldots, p$. After permuting within each block of interventional data across p blocks of $\mathbf{X}_c$, we extract the first n_{1j} rows in $\mathbf{X}_c[\mathscr{I}_{c,j},]$ to form the rows of $\tilde{\mathbf{X}}_m^{(1)}$, and use the remaining n_{2j} rows in the same block of the permuted combined data to form rows in $\tilde{\mathbf{X}}_m^{(2)}$, for $j = 1, \ldots, p$.

Once the pairs bootstrap data sets, $\tilde{\mathbf{X}}_m^{(1)}$ and $\tilde{\mathbf{X}}_m^{(2)}$, are generated, we use $\tilde{\mathbf{X}}_m^{(2)}$ to estimate regression coefficients assuming the structure of $\hat{G}_1$ to obtain unpenalized estimates of $\mathbf{B}$, denoted by $\tilde{\mathbf{B}}_m^{(2)}$. Lastly, evaluating at $\tilde{\mathbf{X}}_m^{(1)}$ and this $\tilde{\mathbf{B}}_m^{(2)}$ yields a bootstrap version of Q_2 under H_0^b. In this pairs bootstrap procedure designed for the scenario where data are from designed experiments, the resampling following permutation is implemented within each experimental condition to preserve the original mixed structure of observational data and interventional data. It is worth pointing out that $\tilde{\mathbf{X}}_m^{(2)}$ obtained from the above pairs bootstrap procedures will not be a sensible analogue of $\mathbf{X}^{(2)}$ if H_0^a is true but H_0^b is not; and thus this procedure cannot be used or easily revised to create a bootstrap analogue of Q_1. Indeed, creating pairs bootstrap data under H_0^a while allowing the possibility that H_0^b is false is much more challenging than creating pairs bootstrap data under H_0^b. We do not pursue this for Q_1 further in the current study.

5.3. Calibration

In assessing the quality of moments estimation for the test statistics in empirical studies, we observe (often severe) underestimation of the variance and (usually mild) underestimation of the mean. These underestimations lead to an inflated size of some tests. The inflation is more noticeable when experimental data are used for inference than when only observational data are available. When it comes to Q_2, the inflation of Type I error is more substantial in the wild bootstrap procedure than in the pairs bootstrap procedure. We conjecture that the phenomenon can be partly explained by reusing $\hat{G}_1$ in all bootstrap samples, and in part due to possible underestimation of error variances $\hat{\sigma}^2_{k,\hat{\mathscr{I}}_{1,j}}$. Re-estimating the graph structure each time new bootstrap data are generated is time-consuming and does not effectively correct the problem, neither does inflating the error variance estimates. We propose to correct for this problem by making two changes in the wild bootstrap procedures depicted in Fig. 3, and similar changes in the pairs bootstrap procedure.

The first change is made to alleviate the concern of reusing $\hat{G}_1$ in all bootstrap samples, without adding too much computational burden. Recall that one needs to choose a tuning parameter λ in Step 2 of the parent selection algorithm in Section 3.1. More specifically, we choose λ from a geometric sequence of K candidate values, $\lambda_{\max} > q\lambda_{\max} > q^2\lambda_{\max} > \cdots > q^{K-1}\text{K-1}\lambda_{\max}$, starting from $\lambda_{\max}$ that leads to a nearly empty graph, then gradually dropping the penalty with $q \in (0, 1)$ such that $q^{K-1}\text{K-1}\lambda_{\max}$ gives a fairly dense graph. Suppose λ^* is the chosen tuning parameter based on SIC using data $\mathbf{X}^{(1)}$. After a bootstrap analogue of $\mathbf{X}^{(1)}$ of the same size, i.e., $\tilde{\mathbf{X}}_m^{(1)}$, is generated, we re-estimate G_1 using a tuning parameter equal to λ^*/q^{r_1}, where r_1 is a pre-specified non-negative integer (to be explained next). In other words, G_1 is re-estimated based on $\tilde{\mathbf{X}}_m^{(1)}$ using a pre-specified value of λ that is no smaller than λ^*. This avoids the most time-consuming part of graphical structure learning, which is the tuning parameter selection. Denote by $\hat{G}_{1,m}$ the resultant estimate of G_1. Then the bootstrap versions of the test statistics, that is, $\tilde{Q}_{1,m}$ and $\tilde{Q}_{2,m}$, are computed assuming the structure of $\hat{G}_{1,m}$. These modified bootstrap versions of test statistics yield much improved mean estimation and slightly improved variance estimation.

The second change is made in the bootstrap sample size in order to further improve the variance estimation of a test statistic. Take the wild bootstrap procedure for Q_1 based on experimental data as an example. Besides the bootstrap procedure depicted in Fig. 3-(a) but with the added step of obtaining $\hat{G}_{1,m}$ after $\tilde{\mathbf{X}}_m^{(1)}$ is generated, we repeat this entire wild bootstrap procedure using a subset of $\mathbf{X}^{(2)}$, denoted by $\mathbf{X}_*^{(2)}$, that consists of $n_{2j}^* = \max(\lfloor r_2 n_{2j} \rfloor, 3)$ rows randomly selected from $\mathbf{X}^{(2)}[\mathscr{I}_{2j},]$, for $j = 1, \ldots, p$, where $r_2 \in (0, 1)$ is a pre-specified quantity to be tuned via simulated data (to be explained next). Denote by $\tilde{Q}_{1,m}^*$ the counterpart of $\tilde{Q}_{1,m}$ resulting from applying the bootstrap procedure in Fig. 3-(a) to $\mathbf{X}_*^{(2)}$. We then use the sample mean of $\{\tilde{Q}_{1,m}\}_{m=1}^M$ and the sample variance of $\{\tilde{Q}_{1,m}^*\}_{m=1}^M$ as the two moments in the method of moments for estimating a gamma distribution as the null distribution of Q_1. When only observational data are available, we choose $\lfloor r_2 N_k \rfloor$ rows of $\mathbf{X}^{(k)}$ randomly to create a subsample $\mathbf{X}_*^{(k)}$ to begin a bootstrap procedure, for $k = 1, 2$. Shao (1996) and references therein discussed the idea of using a smaller sample size in bootstrap as a remedy when the regular bootstrap procedure (with the bootstrap sample size equal to the original sample size) performs unsatisfactorily in the context of model selection. As pointed out in these earlier works, a major hurdle in this remedy is that how much smaller the bootstrap sample size should go, determined by r_2, depends on unknown true model settings. We face a similar question relating to the first change above, that is, the amount of stepping-up from λ^* to λ^*/q^{r_1} when re-estimating G_1 based on bootstrap data, determined by r_1. In what follows, we provide a simulation-based calibration method to choose r_1 and r_2.

After obtaining $\hat{G}_1$ and $\tilde{\mathbf{B}}^{(1)}$ based on $\mathbf{X}^{(1)}$, we treat the estimated model as the true model shared by the two populations. Then, at a chosen combination of (r_1, r_2), for $d = 1, \ldots, D$, we generate two samples, $\mathbf{X}_d^{(1)}$ and $\mathbf{X}_d^{(2)}$, according to this assumed "true" model, where $\mathbf{X}_d^{(k)}$ is of the same size and structure as those of $\mathbf{X}^{(k)}$, for $k = 1, 2$. A proposed testing procedure is carried out based on $\mathbf{X}_d^{(1)}$ and $\mathbf{X}_d^{(2)}$, reaching to a conclusion of rejecting or failing to reject a null hypothesis,

for $d = 1, \ldots, D$, which gives an empirical size of a proposed test. We then choose (r_1, r_2) so that this empirical size across D simulated pairs of samples is slightly below the nominal Type I error. For one data application, this calibration procedure that involves trial-and-error can be implemented very quickly by using a moderate D (e.g, $D = 100$); and an unsatisfactory attempt/guess (for values of r_1 and r_2) typically offers good clues leading to a more satisfactory outcome in a small number of trials. Aiming at an empirical size slightly lower than the nominal level is to avoid an inflated Type I error when using the final chosen (r_1, r_2) to the original data, $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, for a testing procedure.

To improve the stability of the calibration procedure and avoid an overly conservative choice of (r_1, r_2), we recommend using the sample that has richer information for graphical structure learning to obtain an estimated graph. In our notations, this means that, when two samples are of different sizes, we always refer to the bigger data set as $\mathbf{X}^{(1)}$. We implemented the proposed calibration method in a simulation study, where 300 realizations of pairs of samples $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ are generated according to each of twenty graphs shared by the two populations. Across all Monte Carlo replicates, at the nominal level of 0.05, the rejection rate of a testing procedure using the values of (r_1, r_2) chosen by this calibration method, where we aim at an empirical size of 0.03, is never above 0.05, although 20% of these replicates return a rejection rate below 0.03.

6. Numerical study

6.1. Simulation settings

In the simulation experiment, we set two levels for the number of nodes in each graph, $p = 10$ and 20, four levels for the overall sample sizes $N = N_1 = N_2 = t \times p$, with $t = 5, 10, 15, 20$, and two scenarios in regard to whether or not interventional data are available. In the first scenarios, data are from designed experiments, where the number of interventional data points per node is $n_{kj} = t$, for $k = 1, 2$ and $j = 1, \ldots, p$. In the second scenario, data are from observational studies and thus no interventional data are available. When creating the underlying graph structures associated with two populations, G_1 and G_2, we first randomly create a topological ordering of the p nodes. Then, following the ordering, we randomly create $2p$ edges for each acyclic graph, with each node having at most four parents. One only needs to generate one graph structure in a Monte Carlo replicate when generating data under H_0^a or H_0^b.

Once G_1 and G_2 are created, we generate data according to (1) for each node following the topological ordering. More specifically, the model errors $\epsilon_j^{(k)}$ are generated from $N(0, 1)$; and, under the first scenario, interventional data associated with $X_j^{(k)}$ are generated from $N(0, 1)$, independent of model errors, for $k = 1, 2$ and $j = 1, \ldots, p$. Non-zero entries in $\mathbf{B}^{(1)}$ are generated from uniform(0.5, 2), and we consider three settings for $\mathbf{B}^{(2)}$:

(I) $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, corresponding to H_0^b that also implies $H_0^a : G_1 = G_2$;
(II) $\mathbf{B}^{(2)}$ has non-zero entries generated from uniform(0.5, 2), independent of the random numbers generated for non-zero entries in $\mathbf{B}^{(1)}$, with $G_1 \neq G_2$, which creates a scenario under H_1^a that also implies H_1^b;
(III) non-zero entries in $\mathbf{B}^{(2)}$ are generated from uniform(0.5, 2), independent of the random numbers generated for non-zero entries in $\mathbf{B}^{(1)}$, while $G_1 = G_2$, corresponding to the scenario where H_0^a and H_1^b hold simultaneously.

Fig. 4 provides QQ plots of Q_2 based on 300 realizations of Q_2 computed using data generated according to a given graphical configuration under (I). The QQ plots of Q_1 collected under the same scenario are similar to those in Fig. 4. Fig. 5 contains QQ plots of Q_1 based on 300 data sets generated according to a given network configuration under (III). All QQ plots suggest that the null distributions of the proposed test statistics can be well approximated by gamma distributions. Both Figs. 4 and 5 are obtained under the first type of study that produces interventional data besides observational data for each node. We observe similar QQ plots for the test statistics under H_0^a and H_0^b when only observational data are used.

6.2. Simulation results

Under each simulation setting, across 300 Monte Carlo replicates, we monitor how often each test statistic, pairing with a bootstrap procedure with $M = 300$, rejects of a null hypothesis at the nominal level of 0.05. Fig. 6 presents these rejection rates associated with Q_1 and Q_2 when data are from designed experiments.

As seen in Fig. 6-(a) where H_0^b is true, the size associated with Q_2 is close to or slightly lower than the nominal level. The lower size can be due to the adjustment in the bootstrap sample size. The testing procedures based on Q_2 outperform that based on Q_1 in detecting violation of H_0^a. Regardless, designed for detecting differences in the graph structure only, Q_1 has a promising power when $G_1 \neq G_2$ as seen in panel (b) of Fig. 6, although its Type I error rate is slightly inflated when $G_1 = G_2$ but $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ as seen in panel (c), especially when the graphs are sparser (with $p = 20$).

Fig. 7 provides counterpart results when data are from observational studies. In this case, all testing procedures retain a size close to the nominal level. In both types of study presented in Figs. 6 and 7, all testing procedures have a higher power to detect violation of a null hypothesis when the graph is sparser. With $2p$ edges in a graph in the simulation, the graph with a larger p results in a sparser graph.

Although Q_1 and Q_2 are designed for testing different null hypotheses, combining inference results from both can shed more light on the underlying truth. For example, a non-significant Q_1 in conjunction with a significant Q_2 provide evidence from data suggesting that the two populations share the same graphical structure, although associations between some

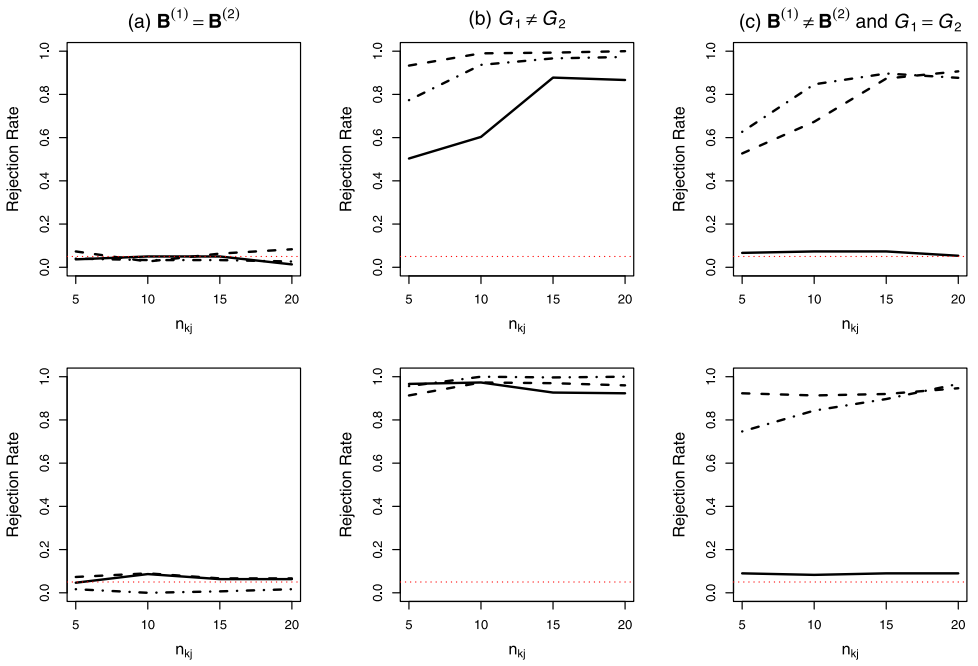


Fig. 6. Rejection rates associated with Q_1 and Q_2 when $p = 10$ (upper panels), 20 (lower panels) across 300 Monte Carlo replicates versus the number of interventional data points per node, n_{kj}, under three settings: (a) $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, (b) $G_1 \neq G_2$, and (c) $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ while $G_1 = G_2$. Each panel contains three sets of rejection rates as n_{kj} varies: rejection rates of Q_1 (solid line), those of Q_2 when wild bootstrap is used to estimate the p-value (dashed line), and those of Q_2 when pairs bootstrap is implemented to estimate the p-value (dash-dotted line). The red dotted line in each panel is the reference line highlighting the nominal level 0.05.

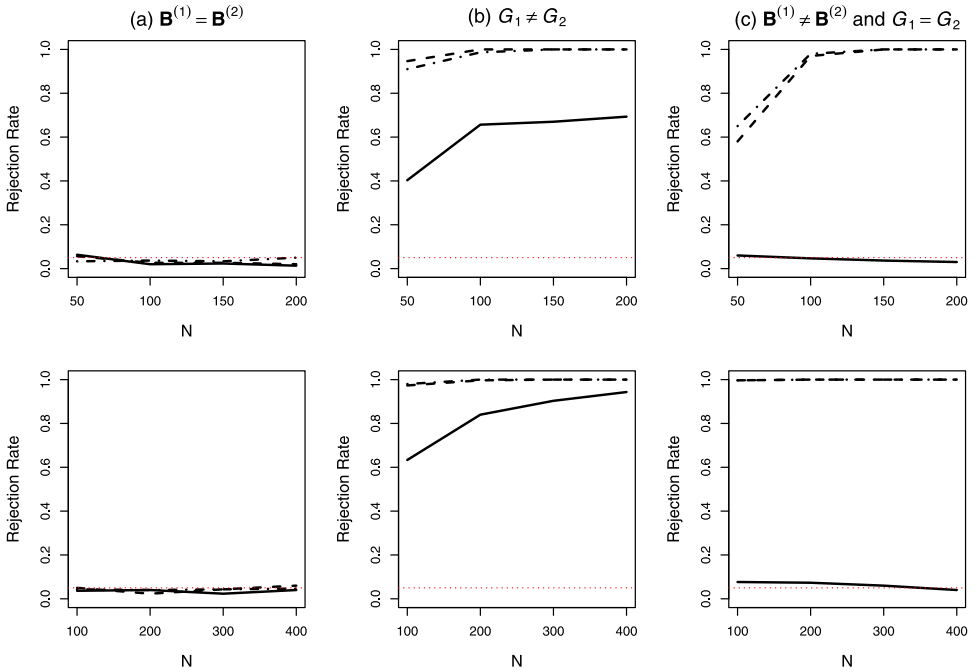


Fig. 7. Rejection rates associated with Q_1 and Q_2 based on data from observational studies when $p = 10$ (upper panels), 20 (lower panels) across 300 Monte Carlo replicates versus the sample size N (common between two samples) under three settings: (a) $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, (b) $G_1 \neq G_2$, and (c) $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ while $G_1 = G_2$. Each panel contains three sets of rejection rates as N varies: rejection rates of Q_1 (solid line), those of Q_2 when wild bootstrap is used to estimate the p-value (dashed line), and those of Q_2 when pairs bootstrap is implemented to estimate the p-value (dash-dotted line). The red dotted line in each panel is the reference line highlighting the nominal level 0.05.

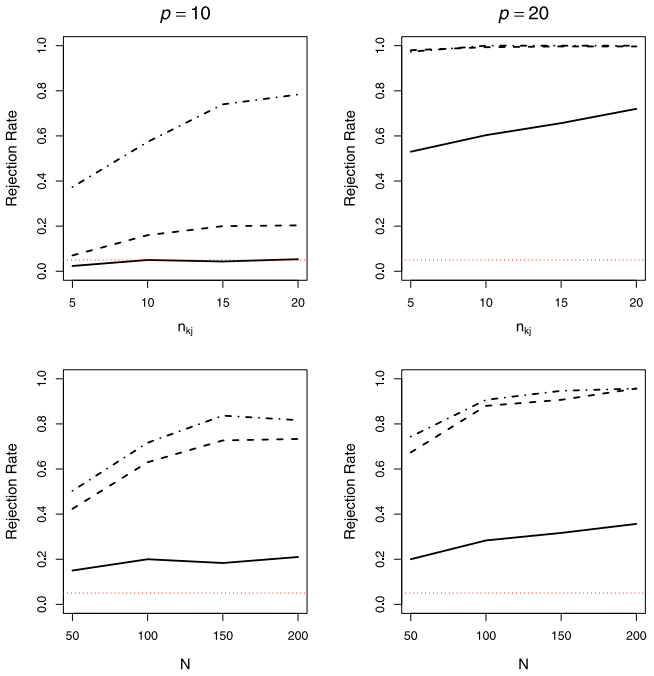


Fig. 8. Rejection rates associated with Q_1 and Q_2 when $p = 10$ (left panels), 20 (right panels) across 300 Monte Carlo replicates versus the number of interventional data points per node, n_{kj}, when experimental data are available (upper panels), or versus the sample size N when only observational data are available (lower panels) under the true-model configurations where G_2 results from randomly deleting 20% of the edges in G_1. Each panel contains three sets of rejection rates as n_{kj} or N varies: rejection rates of Q_1 (solid line), those of Q_2 when wild bootstrap is used to estimate the p-value (dashed line), and those of Q_2 when pairs bootstrap is implemented to estimate the p-value (dash-dotted line). The red dotted line in each panel is the reference line highlighting the nominal level 0.05.

nodes can differ in strength or direction between the two populations. One reaches a contradictory conclusion only when the testing procedures produce a significant Q_1 yet a non-significant Q_2. Fortunately, this rarely happens according to our extensive empirical study. In particular, across all ninety six simulation settings presented here (when varying p, the sample size/structure and true-model configurations), only in nine of them did we observe a relative frequency of reaching to this contradictory conclusion higher than 5% across 300 Monte Carlo replicates. Among these nine settings, the highest relative frequency is merely 8.7%.

6.3. Additional results

In Setting (II) described in Section 6.1, we randomly create two graph structures, which usually lead to G_1 and G_2 very different from each other. Similar to this setting but with less drastically different G_1 and G_2, we add an additional simulation where, after generating G_1, we create G_2 by randomly deleting 20% of the edges in G_1. Fig. 8 shows the empirical power of Q_1 and Q_2 based on different types of data under this setting. With the two populations differ less dramatically in their graph structures, Q_1 rejects H_0^a much less frequently than what are observed under (II) in Figs. 6 and 7, even more so when p is smaller, leading to two structures different in an even smaller number of edges. The empirical power of Q_2 is still promising, although also lower than before as one would expect when violation of the null is less severe than that under (II).

In the algorithm for inferring a Bayesian network described in Section 3.1, we adopt a score-based model selection method developed in an earlier work (Huang and Zhang, 2013) in Step 2. This method entails selecting parents of one node at a time. We thus call it the node-wise parent selection (NPS) algorithm. Other model selection methods can be employed in this step in place of NPS. We conjecture that operating characteristics of Q_1 and Q_2 should not change too much when a different model selection algorithm is used there, as long as one follows a consistent model selection. To confirm this conjecture, we repeated part of the simulation presented in Section 6.2, using the pairwise coordinate descent algorithm (PCD) proposed by Fu and Zhou (2013) in Step 2. Fig. 9 summarizes operating characteristics of the proposed testing procedures observed in this repeated experiment when $p = 10$ based on experimental data. For ease of comparison, we duplicate in Fig. 9 the top panel of Fig. 6, which are for results under the same setting but NPS is used for network estimation.

These results confirm our conjecture that using a different consistent model selection method does not alter too much the phenomena observed in Section 6.2. It is only under Setting (III) one sees substantial discrepancy in the power of Q_2

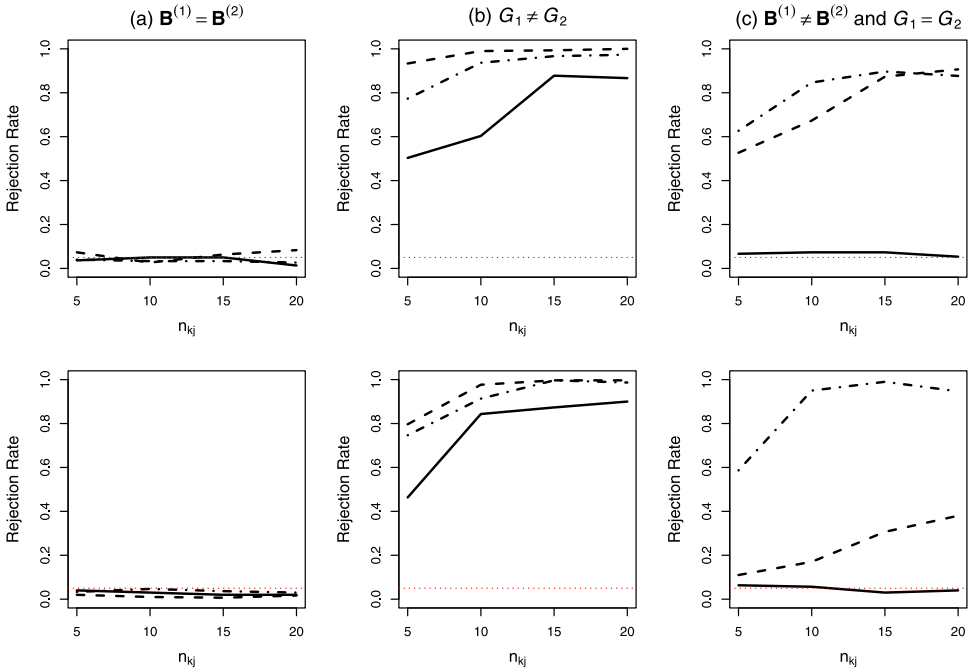


Fig. 9. Rejection rates associated with Q_1 and Q_2 when $p = 10$ with $\hat{G}_1$ obtained via NPS (upper panels) and those via PCD (lower panels) across 300 Monte Carlo replicates versus the number of interventional data points per node, n_{kj}, under three settings: (a) $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, (b) $G_1 \neq G_2$, and (c) $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ while $G_1 = G_2$. Each panel contains three sets of rejection rates as n_{kj} varies: rejection rates of Q_1 (solid line), those of Q_2 when wild bootstrap is used to estimate the p-value (dashed line), and those of Q_2 when pairs bootstrap is implemented to estimate the p-value (dash-dotted line). The red dotted line in each panel is the reference line highlighting the nominal level 0.05.

when the wild bootstrap is used to estimate its p-value. We believe that this big gap can be mostly explained by different amounts of calibration in the wild bootstrap. If an inconsistent model selection method is used, leading to an inaccurate $\hat{G}_1$, the calibration procedure can still guard against size inflation of a test by design. However, the amount of calibration needed to avoid an inflated Type I error will result in a substantial loss of power to reject a null hypothesis.

Lastly, a referee suggested yet another sensible test statistic for testing H_0^b given by

$$\tilde{Q}_2 = \sum_{j=1}^{p} \left\{ n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(1)}) \right\}^{\mathrm{T}} \left\{ \tilde{\mathbf{H}}_j^*(\tilde{\mathbf{B}}_j^{(1)}) \right\}^{-1} \left\{ n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(1)}) \right\},$$

where $\tilde{\mathbf{H}}_j^*(\tilde{\mathbf{B}}_j^{(1)}) = n_{-2j}^{-1} \sum_{\ell \in \mathscr{O}_{2j}} \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(1)}) \Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(1)})^{\mathrm{T}}$, in which $\Psi_{j\ell}(\tilde{\mathbf{B}}_j^{(1)})$ is the normal score evaluated at $\tilde{\mathbf{B}}_j^{(1)}$ and $\mathbf{X}^{(2)}[\ell, \,]$, and $\tilde{\mathbf{B}}_j^{(1)}$ is unpenalized estimate of $\mathbf{B}$ assuming the structure of $\hat{G}_1$. The wild bootstrap and pair bootstrap procedure developed in Section 5 can be easily revised to approximate p-values associated with $\tilde{Q}_2$. To better contrast $\tilde{Q}_2$ with Q_2, we re-express $\tilde{Q}_2$ as $\tilde{Q}_2(\hat{G}_1, \tilde{\mathbf{B}}^{(1)}, \mathbf{X}^{(2)})$, and write Q_2 defined in as $Q_2(\hat{G}_1, \tilde{\mathbf{B}}^{(2)}, \mathbf{X}^{(1)})$. Now one can see that $\tilde{Q}_2$ differs from Q_2 first in that it involves estimating $\mathbf{B}^{(1)}$ while assuming the structure of $\hat{G}_1$, whereas Q_2 requires estimating $\mathbf{B}^{(2)}$ assuming the structure of $\hat{G}_1$. If $G_1 = G_2$, then $\hat{G}_1$ also serves as an estimate for G_2, and thus this first difference is not expected to cause noticeable discrepancy between $\tilde{Q}_2$ and Q_2. If $G_1 \neq G_2$, by estimating $\mathbf{B}^{(2)}$ assuming a misleading estimated structure specified by $\hat{G}_1$, Q_2 has the potential to magnify the disagreement between two samples when evaluating it at $\tilde{\mathbf{B}}^{(2)}$ and $\mathbf{X}^{(1)}$.

Under simulation settings described in Section 6.1, we compare in Fig. 10 empirical power of $\tilde{Q}_2$ with that of Q_2 when $p = 10$ based on experimental data. Although both are sensible test statistics for testing H_0^b, Q_2 is at least as powerful as $\tilde{Q}_2$ in the presence of either form of deviation from the null.

7. Real-life data applications

In this section we entertain data examples originating from two applications to test differential Bayesian networks between two populations.

Example 1 (*Flow Cytometry*)**.** In this example, we consider a flow cytometry data set consisting of $p = 11$ phosphomolecular measurements from each of 7466 human immune system cells collected in an experiment described

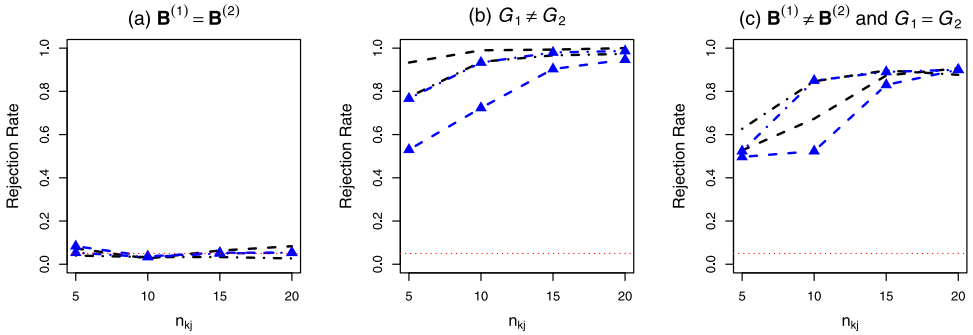


Fig. 10. Rejection rates associated with Q_2 and $\tilde{Q}_2$ when $p = 10$ across 300 Monte Carlo replicates versus the number of interventional data points per node, n_{kj}, under three settings: (a) $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$, (b) $G_1 \neq G_2$, and (c) $\mathbf{B}^{(1)} \neq \mathbf{B}^{(2)}$ while $G_1 = G_2$. Each panel contains four sets of rejection rates as n_{kj} varies: rejection rates of Q_2 when wild bootstrap is used to estimate the p-value (dashed line), those of Q_2 when pairs bootstrap is implemented to estimate the p-value (dash-dotted line), and the counterpart rejection rates of $\tilde{Q}_2$ depicted using blue lines of the same line types as above that run through solid triangles. The red dotted line in each panel is the reference line highlighting the nominal level 0.05. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

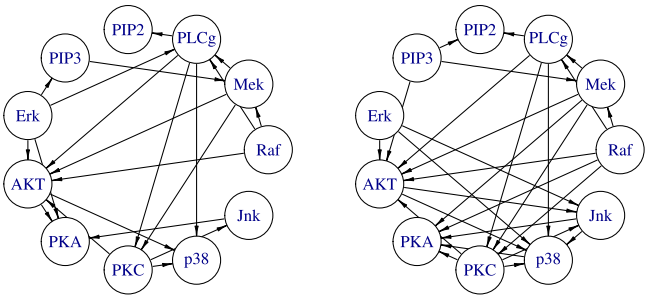


Fig. 11. The estimated graph structures based on the two data sets obtained from (nearly) equal split of data within each experimental condition across the nine conditions that yield the flow cytometry data.

in Sachs et al. (2005). In this experiment, a series of stimulatory cues and inhibitory interventions were imposed to create nine experimental conditions described in Table 1 in Sachs et al. (2005), under which experimental data for the eleven phosphorylated proteins and phospholipids are collected. Fu and Zhou (2013) applied a likelihood-based penalized estimation method on the entire data set to infer one directed signaling network. Peterson et al. (2015) viewed this data set consisting of nine samples coming from nine populations, treating data from one experimental condition as a sample from one population, and inferred one undirected graph using each of the nine samples.

To illustrate an application of the testing procedures in a scenario under the null, we randomly split the data collected under each experimental condition into two halves of equal or nearly equal size, then half of data under each condition goes in $\mathbf{X}^{(1)}$ and the other half contributes to $\mathbf{X}^{(2)}$. This produces two data sets, each of size $N_1 = N_2 = 3733$, which can be reasonably assumed to arise from some common underlying populations, creating a scenario under $H_0^b : \mathbf{B}^{(1)} = \mathbf{B}^{(2)}$. Based on Q_1 computed from the so-obtained two samples, and using the wild bootstrap procedure proposed in Section 5.1 for experimental data, we fail to reject $H_0^a : G_1 = G_2$ with an estimated p-value of 0.991 using 300 bootstrap samples. We also computed Q_2, and applied the wild bootstrap and the pairs bootstrap to estimate its null distribution, and fail to reject H_0^b, with estimated p-values being 0.964 and 0.120, respectively. The fact that Q_2 fails to reject H_0^b can also be interpreted as data evidence in favor of H_0^a since $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$ implies $G_1 = G_2$. Fig. 11 presents the two estimated graph structures. Even though the graph in the right panel, $\hat{G}_2$, is denser than the one in the left panel, $\hat{G}_1$, the additional edges in $\hat{G}_2$ are much less significant than other edges that also exist in $\hat{G}_1$. In particular, among the fifteen edges in $\hat{G}_2$ but not in $\hat{G}_1$, eleven of them correspond to estimated regression coefficients whose p-values are above 10^{-2}. All but three edges in $\hat{G}_1$ have p-values below 10^{-4}, so are most edges in $\hat{G}_2$ that are also in $\hat{G}_1$. This can be where the tuning parameter chosen based on $\mathbf{X}^{(2)}$ is not large enough to lead to a more parsimonious model. Despite these discrepancies in the selected models based on $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, the proposed testing procedures are able to take into account the uncertainty in graph estimation and lead to the correct conclusion that the current Bayesian network of eleven nodes is not differential between the two (made-up) populations.

Example 2 (*Bile Acids*)**.** There have been increasing evidence suggesting that biodirectional biochemical communication between the brain and the gut contributes to a variety of neurodegenerative diseases, such as Alzheimer's disease

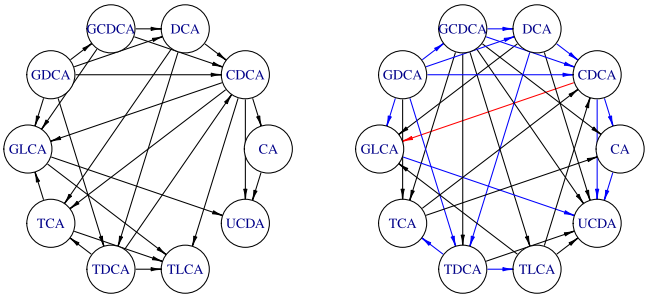


Fig. 12. The estimated network based on bile acids data from cognitively normal subjects (on the left) and the one based on data from subjects with Alzheimer's disease (on the right). Edges in the AD graph that are also in the CN graph are highlighted in blue. The edge in the AD graph of which the estimated regression coefficient is of the opposite sign as the counterpart estimated regression coefficient in the CD graph is highlighted in red (pointing from CDCA to GLCA). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(AD) (Mohajeri et al., 2018; Ambrosini et al., 2019; Ma et al., 2019; Mahmoudian Dehkordi et al., 2019). In this example, we analyze levels of $p = 10$ primary bile acids in human serum samples of two groups of individuals who were recruited in Alzheimer's Disease Neuroimaging Initiative (http://adni.loni.usc.edu) Phase 2 study. The two samples used to infer the two Bayesian networks of the ten bile acids are a sample of $N_1 = 182$ cognitively normal (CN) subjects and a sample of $N_2 = 132$ AD subjects, respectively. The goal is to test, using data from an observational study, whether or not the Bayesian network of these ten bile acids is differential between the CN population and the AD population.

Applying the graph estimation algorithm in Section 3 to each of the two samples gives the two estimated graphs shown in Fig. 12. We then computed Q_1 and Q_2 and estimated their p-values, obtaining estimated p-values of 0.390 for Q_1 and 0.032 for Q_2 when using the pairs bootstrap. We interpret these as evidence for potentially similar graph structures between the two networks but more substantial difference in associations between some bile acids across the two populations. Looking more closely at the two estimated networks in Fig. 12, one can see that two graphs are similar in structure, sharing many common directed edges. In particular, the two graphs have the same edges connecting DCA, GDCA, and TDCA, and the corresponding estimated regression coefficients are of the same sign in $\hat{\mathbf{B}}^{(1)}$ and $\hat{\mathbf{B}}^{(2)}$. These three bile acids are among the key members along the bile acids metabolism primary pathway (see, e.g., Figure 3 in Mahmoudian Dehkordi et al., 2019). In contrast, the two graphs show different patterns of connections between CDCD, GLCA, and TLCA, the three bile acids in the alternative pathway of metabolism. These connected edges are either of opposite directions (e.g., the edge connecting GLCA and TLCA) between the two graphs, or the corresponding estimated regression coefficients are of opposite signs (for the edge pointing from CDCA to GLCA).

These statistical evidence suggest that the overall signaling pattern and functionality of the bile acids may be similar between the CN population and AD population, especially when the primary pathway of metabolism is concerned. But the strength and direction of associations of bile acids along the alternative pathway may be altered due to AD, even though the alternation is not significant enough to create an overall discrepancy in the two graph structures to be detected by Q_1.

8. Discussion

Aiming to compare two Bayesian networks using data either from observational studies or designed experiments, we propose two test statistics based on quadratic inference functions associated with the normal score function. The first proposed test statistic Q_1 can tell apart two structurally different networks, and does not distinguish between two networks of the same structure with different strengths of associations between nodes across two populations. Consequently, the proposed testing procedure based on Q_1 can be useful when connectivity between nodes is the focal point of a study. The second proposed test statistic, Q_2, serves as a sensitive indicator of the difference between the two networks, whether the difference lies in the structure or the strength of association, or in both regards. An appealing feature of Q_1 and Q_2 is that one only needs to estimate one graph structure G_1 to compute them, and thus bypasses the daunting task of learning multiple graph structures. Because this graph structure learning is the first step of the testing procedures, it is recommended to set the sample that contains richer information for structure learning to be $\mathbf{X}^{(1)}$. This is consistent with the recommendation given in the calibration method in Section 5.3.

Motivated by a referee's question relating to which sample is viewed as $\mathbf{X}^{(1)}$ in our notations, we envision alternative test statistics. Take testing H_0^a as an example which is based on Q_1. After one computes Q_1 based on a given ordering (in terms of which sample is referred to $\mathbf{X}^{(1)}$ and which is called $\mathbf{X}^{(2)}$), one flips the ordering and computes Q_1 again. Then one uses a weighted average of these two versions of Q_1 to test H_0^a. With an adequate choice of the weights, this strategy can yield a statistic closer to a distance measure between two graphs that is invariant to which sample is viewed

as $\mathbf{X}^{(1)}$. Bootstrap procedures for estimating the null distributions of the so-constructed test statistics require more careful development though.

Thanks to the regression framework employed in this line of study, we believe that these testing procedures can be easily generalized to test differential non-Gaussian Bayesian networks, where regression models with non-Gaussian model errors are used to characterize a Bayesian network. Moreover, other score functions adapted to a regression setting can be adopted in place of the normal score function to construct new test statistics based on quadratic inference functions that are robust to normality assumption or other parametric assumptions imposed on a Bayesian network. On the computational side, the current test statistics can be easily computed for networks with a small or moderate number of nodes p, but the computation can be challenging when p is large due to their dependence on the inverse of a $p \times p$ matrix. To develop new testing procedures that are more scalable for large graphs is among our follow-up research goals.

Besides allowing a larger p, another future research direction is to consider more than two populations, say, $K(> 2)$ of them. Even though in this case one may carry out pairwise comparison based on the current testing procedures, one needs to adjust for multiple testing, which is not always trivial. Adopting notations introduced in Section 6.3, for testing $H_0^b : \mathbf{B}^{(1)} = \cdots = \mathbf{B}^{(K)}$, a sensible test statistic can be in the form of $\max_{1\le k\le K} Q_{2,k}(\hat{G}_c, \hat{\mathbf{B}}_c, \mathbf{X}^{(k)})$, where $\mathbf{X}^{(k)}$ is the sample from population k, for $k = 1, \ldots, K$, $\hat{G}_c$ and $\hat{\mathbf{B}}_c$ are the estimated graph structure and regression coefficients matrix obtained using information from all K samples, assuming the null is true. The first challenge in this line of development is to obtain sensible $\hat{G}_c$ and $\hat{\mathbf{B}}_c$.

Even though the proposed tests are expected to distinguish between two equivalent classes of networks, as opposed to two networks, when only observational data are available, we did not always observe in the simulation study compromise in power of the proposed testing procedures when applied to observational data only. Operating characteristics of these tests based on data that present network identifiability complications deserve more systematic investigation. When some causal relationships can be identified owing to certain designs of interventions, new test statistics that incorporate inference for causality can lead to more powerful tests. The current tests are developed by considering inference for association but not causality. When a null hypothesis is rejected by them, a natural follow-up task is to zoom in on a small collection of nodes between which causality relationships have high power to differentiate two populations. We have started tackling this problem borrowing ideas relating to causal inference in Peters et al. (2016).

Acknowledgments

The authors wish to thank the Editor-in-Chief, anonymous Associate Editor, and two referees for their insightful comments and suggestions that greatly improved the manuscript. Zhang's work is partially supported by National Institutes of Health, NIAID, USA grant R01AI121226.

References

Ambrosini, Y.M., Borcherding, D.C., Kanthasamy, A., Kim, H.J., Willette, A.A., Jergens, A.E., Allenspach, K., Mochel, J.P., 2019. The gut-brain-axis in neurodegenerative diseases and relevance of the canine model: A review. Front. Aging Neurosci. 11, 130.

Andersson, S.A., Madigan, D., Perlman, M.D., 1997. A characterization of Markov equivalence classes for acyclic digraphs. Ann. Statist. 25 (2), 505–541.

Aragam, B., Zhou, Q., 2015. Concave penalized estimation of sparse Gaussian Bayesian networks. J. Mach. Learn. Res. 16 (1), 2273–2328.

Chickering, D.M., 2002. Learning equivalence classes of Bayesian-network structures. J. Mach. Learn. Res. 2 (Feb), 445–498.

Chung, F., Chung, F.R., Graham, F.C., Lu, L., Chung, K.F., et al., 2006. Complex Graphs and Networks, No. 107. American Mathematical Society.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2001. Introduction to Algorithms, second ed. The MIT Press.

Durante, D., Dunson, D.B., et al., 2018. Bayesian inference and testing of group differences in brain networks. Bayesian Anal. 13 (1), 29–58.

Edwards, D., 2012. Introduction to Graphical Modelling. Springer Science & Business Media.

Ellis, B., Wong, W.H., 2008. Learning causal Bayesian network structures from experimental data. J. Amer. Statist. Assoc. 103 (482), 778–789.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.

Freedman, D.A., et al., 1981. Bootstrapping regression models. Ann. Statist. 9 (6), 1218–1228.

Fu, F., Zhou, Q., 2013. Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. J. Amer. Statist. Assoc. 108 (501), 288–300.

Gill, R., Datta, S., Datta, S., 2010. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 11 (1), 95.

Hauser, A., Bühlmann, P., 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. J. Mach. Learn. Res. 13 (Aug), 2409–2464.

Huang, X., Zhang, H., 2013. Variable selection in linear measurement error models via penalized score functions. J. Statist. Plann. Inference 143 (12), 2101–2111.

Huang, X., Zhang, H., 2021. Corrected score method for estimating directed acyclic graphs with error-prone nodes. Stat. Med. http://dx.doi.org/10.1002/sim.8925.

Jacob, L., Neuvial, P., Dudoit, S., et al., 2012. More power via graph-structured tests for differential expression of gene networks. Ann. Appl. Stat. 6 (2), 561–600.

Jensen, F.V., 1996. An Introduction to Bayesian Networks, Vol. 210. UCL press London.

Kahn, A.B., 1962. Topological sorting of large networks. Commun. ACM 5 (11), 558–562.

Lauritzen, S.L., 1996. Graphical Models, Vol. 17. Clarendon Press.

Lindsay, B.G., Qu, A., 2003. Inference functions and quadratic score tests. Statist. Sci. 394–410.

Ma, Q., Xing, C., Long, W., Wang, H.Y., Liu, Q., Wang, R.-F., 2019. Impact of microbiota on central nervous system and neurological diseases: the gut-brain axis. J. Neuroinflammation 16 (1), 53.

Mahmoudian Dehkordi, S., Arnold, M., Nho, K., Ahmad, S., Jia, W., Xie, G., Louie, G., Kueider-Paisley, A., Moseley, M.A., Thompson, J.W., et al., 2019. Altered bile acid profile associates with cognitive impairment in Alzheimer's disease—an emerging role for gut microbiome. Alzheimer's Dement. 15 (1), 76–92.

Mohajeri, M.H., La Fata, G., Steinert, R.E., Weber, P., 2018. Relationship between the gut microbiome and brain function. Nutr. Rev. 76 (7), 481–496.
Neapolitan, R.E., 2012. Probabilistic Reasoning in Expert Systems: Theory and Algorithms. CreateSpace Independent Publishing Platform.
Nielsen, T.D., Jensen, F.V., 2009. Bayesian Networks and Decision Graphs. Springer Science & Business Media.
Pearl, J., 2014. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.
Peters, J., Bühlmann, P., Meinshausen, N., 2016. Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. Ser. B Stat. Methodol..
Peterson, C., Stingo, F.C., Vannucci, M., 2015. Bayesian inference of multiple Gaussian graphical models. J. Amer. Statist. Assoc. 110 (509), 159–174.
Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science 308 (5721), 523–529.
Shao, J., 1996. Bootstrap model selection. J. Amer. Statist. Assoc. 91 (434), 655–665.
Städler, N., Dondelinger, F., Hill, S.M., Akbani, R., Lu, Y., Mills, G.B., Mukherjee, S., 2017. Molecular heterogeneity at the network level: high-dimensional testing, clustering and a tcga case study. Bioinformatics 33 (18), 2890–2896.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 267–288.
Verma, T., Pearl, J., 1991. Equivalence and Synthesis of Causal Models. UCLA, Computer Science Department.
Wu, C.-F.J., 1986. Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. 1261–1295.
Xia, Y., Cai, T., Cai, T.T., 2015. Testing differential networks with applications to the detection of gene-gene interactions. Biometrika 102 (2), 247–266.
Zhao, S.D., Cai, T.T., Li, H., 2014. Direct estimation of differential networks. Biometrika 101 (2), 253–268.
Zhao, S., Ottinger, S., Peck, S., Mac Donald, C., Shojaie, A., 2019. Network differential connectivity analysis. arXiv preprint arXiv:1909.13464.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.