Computational Statistics & Data Analysis, Volume 157, May 2021, Article number 107156
DOI:10.1016/j.csda.2020.107156

Gaussian Bayesian network comparisons with graph ordering unknown

Hongmei Zhang[a,*], Xianzheng Huang[b], Shengtong Han[c], Faisal I. Rezwan[d], Wilfried Karmaus[a], Hasan Arshad[e,f], John W Holloway[g]

[a] *Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA.*
[b] *Department of Statistics, University of South Carolina, Columbia, SC, USA.*
[c] *Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee, WI, USA.*
[d] *School of Water, Energy and Environment, Cranfield University, Cranfield, Bedfordshire, UK.*
[e] *Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Sothampton, UK.*
[f] *David Hide Asthma and Allergy Research Centre, Isle of Wight, UK.*
[g] *Human Development and Health, Faculty of Medicine, University of Southampton, Sothampton, UK.*

Abstract

A Bayesian approach is proposed that unifies Gaussian Bayesian network constructions and comparisons between two networks (identical or differential) for data with graph ordering unknown. When sampling graph ordering, to escape from local maximums, an adjusted single queue equi-energy algorithm is applied. The conditional posterior probability mass function for network differentiation is derived and its asymptotic proposition is theoretically assessed. Simulations are used to demonstrate the approach and compare with existing methods. Based on epigenetic data at a set of DNA methylation sites (CpG sites), the proposed approach is further examined on its ability to detect network differentiations. Findings from theoretical assessment, simulations, and real data applications support the efficacy and efficiency of the proposed method for network comparisons.

Keywords: Bayesian methods, DNA methylation, Single Queue Equi-Energy, Differential Gaussian Bayesian network, Variable selections, Ordering

*Corresponding author
Email address: hzhang6@memphis.edu (Hongmei Zhang)

1. Introduction

1.1. An epigenetic epidemiological study

Our work was motivated by a recent epidemiological study aiming to examine joint activities of some epigenetic sites on certain genes. Epigenetics reflects memories of past exposure or physical changes in life and regulates gene functionalities without changing the DNA sequence. DNA methylation at Cytosine-phosphate-Guanine (CpG) sites is one of the most widely studied epigenetic mechanisms and its role is of particular interest due to its known responsiveness to environmental exposures (Felix et al., 2017; Joubert et al., 2016).

In our epidemiological study, the goal is to find out whether the joint activities of certain CpG sites are different between subjects exposed to in utero smoking and those not exposed. Consequently, if they are different, then what are the possible driving DNA methylation sites? Epigenetic changes due to in utero exposure to smoke have been detected at certain CpG sites (Joubert et al., 2012, 2016). However, existing studies have been focusing on effects of individual CpG sites and joint activities among the sites are completely overlooked. Thus, a novel route to appropriately answer the study questions needs to be explored.

Joint activities among genetic or epigenetic factors are commonly described by networks. In general, two types of networks are commonly applied, directed and undirected networks. To identify potential driving epigenetic factors leading to network differentiation, as the goal in our epigenetic epidemiological study, directed networks are of great interest. To examine the impact of environmental exposures, such as in utero exposure to smoke, on gene activities, differences between networks under different conditions are of greater interest than a particular network. There exist methods to infer multiple directed networks under different conditions, e.g., Wang et al. (2018). However, because of the complexity in the process of learning networks, constructed networks are subject

to large variability. Thus, observed differences in networks under different conditions can be simply due to random variability, leading to false discovery of markers. Rigorously comparing networks under different conditions via statistical testing will potentially reduce such false discoveries. In the next session, we briefly review the literature in network construction and network testing.

1.2. Literature review

Networks can be inferred by use of graphical models. Practically, the inferred networks enable a depiction of concrete connections between different variables. Networks or graphs can be directed, that is, one epigenetic site can be a probabilistic stimulus ("parent node") of the other ("child node"). In our study, the benefit of directed networks is that they allow us to identify potential driving epigenetic sites that potentially cause changes of other sites, a unique property of directed networks. Bayesian networks, also noted as probabilistic directed acyclic graphs (DAGs), are directed networks and DAGs accompanied by probabilistic connections between edges. A graph is a DAG if all the links (edges) have directions, but none of the nodes is directly go to itself or through a path to itself (a circle). Gaussian Bayesian networks are the focus of our work such that the association between parents and a child can be described using linear regressions. Graphs can also be undirected, in which case two nodes are associated but one is not a potential predictor of the other. Some other graphs are the mixture of the two (Andersson et al., 1997; Chickering, 2002). Ni et al. (2018) has a comprehensive summary on definitions of different types of graphs.

In Bayesian networks, a range of studies focus on methods dealing with ordered data (i.e., ordering of graph is known) when constructing networks. An ordering of a graph informs possible "parents" of each node. In many applications, data come with a natural ordering. For instance, in gene transcription process, the direction of information flow (graph ordering) is known. Assuming the ordering is known, Shojaie and Michailidis (2010) proposed an efficient penalized likelihood method to estimate adjacent matrices of directed graphs, and Altomare et al. (2013) proposed an objective method for Bayesian net-

work inference. Cao et al. (2019) suggested a class of priors for the purpose of inferring Bayesian networks for ordered data, and Park and Klabjan (2017) proposed a mixed integer programming model and iterative algorithms based on given topological ordering to infer Bayesian networks. Some other works in this area for Gaussian Bayesian network inferences, such as Ben-David et al. (2011); Consonni et al. (2017), are noted and discussed in Cao et al. (2019, 2020).

In other situations, however, graph ordering is unknown as in our motivating example, or partially known as noted in Rahman et al. (2019). Many algorithms and approaches have been proposed to infer Bayesian networks under such a circumstance, including greedy local search (Heckerman et al., 1995), Optimal Reinsertion search (Moore and Wong, 2003), Max-Min Hill-Climbing (Tsamardinos et al., 2006), genetic algorithm (Larrañaga et al., 1996; Lee et al., 2010), dynamic programming (Eaton and Murphy, 2012), branch-and-bound algorithm (Campos and Ji, 2011), likelihood approach with L_1-penalty (Fu and Zhou, 2013), penalized marginal likelihood approach (Oates et al., 2016), and Markov Chain Monte Carlo (MCMC) approaches (Madigan et al., 1995, 1996; Giudici et al., 1999; Ellis and Wong, 2008; Zhou, 2011; Han et al., 2014; Kuipers and Moffa, 2017). Some other works, e.g., Friedman and Koller (2003); Han et al. (2016), infer Bayesian networks by introducing graph ordering MCMC. Permutations have also been used to infer graphs, e.g., the work by Squires et al. (2020). This type of methods is not sensitive to Gaussian assumptions and thus their applications are not limited to Gaussian Bayesian networks. Some network construction methods can be applied to both ordered or unordered data. One example in this direction is the maximum penalized likelihood algorithm proposed by Li and Zhou (2019). However, when ordering unknown, this approach is not able to infer direction of connections and a constructed network reflects underlying correlations between nodes.

Regardless of the status of ordering, most existing works focus on inferring networks. Effort on network comparisons was relatively limited, especially in the area of Bayesian networks. Gill et al. (2010) proposed a procedure to globally test differential undirected graphs particularly applied to genes, based on

strength of genetic associations or interaction between genes. Jacob et al. (2012) tested multivariate two-sample means on known graphs utilizing Hotelling's T^2-tests. Zhao et al. (2014) developed a method to estimate the differences in precision matrices between two differential undirected networks, which was later extended with the ability to globally test differentiation of undirected graphs (Xia et al., 2015). The work by Städler et al. (2017) was under a similar framework, that is, testing differentiation of undirected graphs based on precision matrices. Methods built upon associations of undirected networks with a feature of interest have been proposed as well (Durante et al., 2018). Undirected graphs focus on associations between nodes and do not have the ability to infer causal-effects relationships. On the other hand, Bayesian networks are suitable for experimental data resulted from causal-effects relationships as well as for observational data such that causal relations are unknown. For network testing, Canonne et al. (2017) discussed approaches to test for identity (whether an estimated Bayesian network is equal to a given network) and for closeness (whether two networks are identical or differential). For both types of testing, their proposed algorithms have a probability of 2/3 to detect the underlying truth. Following our motivating example, we aim to compare Bayesian networks constructed under two different conditions, e.g., exposed or not-exposed to smoking in utero, with respect to network structure, direction of node connection, and strength of connection. Thus, we aim at network construction with ordering unknown as well as network testing for closeness between two inferred graphs. Almudevar (2010) proposed an approach to compare two Bayesian networks based on likelihood ratios. Each graph is constructed using minimum spanning trees and utilizes permutations to calculate an empirical p-value for decision-making. However, this approach assumes joint density of two nodes is at least as large as the multiplication of their individual density, which is a relatively strong assumption, implying a potential impossibility of inferring networks correctly (up to Markov equivalence) if using one group of data. With ordering unknown, approaches that can both construct Bayesian networks and test for differentiation between Bayesian networks are lacking. The work pre-

sented in this article is an attempt to address this gap.

In this article, we propose an approach targeted at data with unknown graph ordering. It has the ability of constructing and statistically comparing Bayesian networks under two conditions. Bayesian network constructions and comparisons for ordered data is a special case of the proposed method. Specifically, we consider data from two populations and present a Bayesian method to build Bayesian networks, and make an inference on whether the two populations share the same network (i.e., an identical network) or the networks are differential. To achieve the goal of efficient network comparison, we investigated the conditional posterior probability mass function for network differentiation and approximated the conditional posterior to ensure efficient convergence. The remaining of the article is organized as follows. We introduce in Section 2 the statistical model, likelihood function, prior distributions, and posterior computing. The property of a penalty-incorporated posterior probability is also discussed in this section. Simulations are discussed in Section 3. We present several real data applications to demonstrate the method in Section 4, and summarize our work in Section 5.

2. Methodology

To infer whether two networks are differential or identical, we start from the definition of Bayesian networks in two populations.

2.1. The model

Let $\boldsymbol{X}_{n_x \times p} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p)$ and $\boldsymbol{Y}_{n_y \times p} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_p)$ denote measures of a set of variables, e.g., DNA methylation levels at a set of CpG sites, in two samples from two populations (e.g., exposed vs. non-exposed to smoke in utero) for the same set of p CpG sites (or p nodes in general), respectively, where n_x and n_y are the numbers of observations with $n = n_x + n_y$.

For a given graph ordering $\mathcal{O}$, conditional on the parents, each node is re-

gressed on its parents as

$$\boldsymbol{X}_j = \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i + \boldsymbol{\epsilon}_j^{(x)} \tag{1}$$

and

$$\boldsymbol{Y}_j = \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i + \boldsymbol{\epsilon}_j^{(y)}, \tag{2}$$

where $\boldsymbol{\epsilon}_j^{(x)}$ and $\boldsymbol{\epsilon}_j^{(y)}$, $j = 2, \cdots, p$, are random noise following normal distributions $\boldsymbol{\epsilon}_j^{(x)} \sim N(\boldsymbol{0}, \sigma_{x(j)}^2 \boldsymbol{I})$ and $\boldsymbol{\epsilon}_j^{(y)} \sim N(\boldsymbol{0}, \sigma_{y(j)}^2 \boldsymbol{I})$, respectively, with $\boldsymbol{I}$ being the identity matrix.

If two networks are identical, then they have the same structure as well as the same strength of connection between nodes. In this case, we have $\beta_{ij}^{(1)} = \beta_{ij}^{(2)} = \beta_{ij}^{(c)}$, and consequently we can combine the data to infer a unified network,

$$\begin{pmatrix} \boldsymbol{X}_j \\ \boldsymbol{Y}_j \end{pmatrix} = \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \begin{pmatrix} \boldsymbol{X}_j \\ \boldsymbol{Y}_j \end{pmatrix} + \boldsymbol{\epsilon}_j^{(c)},$$

where $\boldsymbol{\epsilon}_j^{(c)} = \begin{pmatrix} \boldsymbol{\epsilon}_j^{(x)} \\ \boldsymbol{\epsilon}_j^{(y)} \end{pmatrix}$. On the other hand, if the two networks are differential, then each network has its own structure or its own set of coefficients describing the relations between parents and a child. We denote $\beta_{ij}^{(1)} = \beta_{ij}^{(x)}$ and $\beta_{ij}^{(2)} = \beta_{ij}^{(y)}$ as the set of coefficients based on samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, and $\beta_{ij}^{(x)} \neq \beta_{ij}^{(y)}$ for at least one node j, $j = 1, \cdots, p$.

Linking the above two settings together, we re-define $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ as

$$\beta_{ij}^{(1)} = \eta \beta_{ij}^{(c)} + (1-\eta) \beta_{ij}^{(x)}$$

and

$$\beta_{ij}^{(2)} = \eta \beta_{ij}^{(c)} + (1-\eta) \beta_{ij}^{(y)}, \tag{3}$$

where η is an indicator with $\eta = 1$ denoting that the two networks are identical and parameters $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ carry information on network structures as well

as strength of links between nodes. With the re-defined $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$, (1) and (2) can be generalized to the following,

$$\boldsymbol{Z}_j|(\boldsymbol{Z}_i)_{1\leq i<j} = \sum_{i=1}^{j-1}(\eta\beta_{ij}^{c} + (1-\eta)\beta_{ij}^{(z)})\boldsymbol{Z}_i + \epsilon_j^{(z)},$$

where $i : 1 \leq i \leq j-1$ is the candidate parent set of nodes, $\boldsymbol{Z}_j = \boldsymbol{X}_j$, or $\boldsymbol{Y}_j$, and correspondingly, z is x or y. To infer whether the two networks are identical ($\eta = 1$) or not, we apply a Bayesian method discussed in the next sections.

2.2. Prior distributions

For the parameter η that determines whether two networks are identical, we assume no prior knowledge on its preference and choose Bernoulli $Ber(0.5)$ for its prior distribution. In practice, sparse networks are preferred, defined as $|E_0| = O(p)$ with $|E_0|$ being the number of edges of a graph (Preiss, 2008). To determine the parents of a node, we adopt the concept of variable selection when selecting prior distributions for $\beta_{ij}^{(c)}$, $\beta_{ij}^{(x)}$, and $\beta_{ij}^{(y)}$ in $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$. Various options are available. Here we choose a mixture of a Normal distribution and a point mass, also known as a spike and slab model (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005). Conditional on η and $\mathcal{O}$,

$$\beta_{ij}^{(c)}|\eta = 1, \mathcal{O}, r_{ij}^{(c)} \sim r_{ij}^{(c)}N(0, V_c) + (1 - r_{ij}^{(c)})I_{\{\beta_{ij}^{(c)}=0\}},$$
$$\beta_{ij}^{(x)}|\eta = 0, \mathcal{O}, r_{ij}^{(x)} \sim r_{ij}^{(x)}N(0, V_x) + (1 - r_{ij}^{(x)})I_{\{\beta_{ij}^{(x)}=0\}}$$

and

$$\beta_{ij}^{(y)}|\eta = 0, \mathcal{O}, r_{ij}^{(y)} \sim r_{ij}^{(y)}N(0, V_y) + (1 - r_{ij}^{(y)})I_{\{\beta_{ij}^{(y)}=0\}},$$

where $r_{ij}^{(c)}, r_{ij}^{(x)}, r_{ij}^{(y)}$ are indicators denoting the inclusion of node i as a parent of node j, $i = 1, 2, \cdots, j-1$, in a given graph ordering $\mathcal{O}$. Note that the same ordering between the two populations is assumed. This assumption is driven by the motivation example of different joint activities of DNA methylation sites due to different exposures such as in utero exposure to smoke. Given the functionality of genes and epigenetic sites, at least the ordering between the two

populations is expected to be the same to ensure meaningful underlying biological mechanism for the same species (although each population has its unique feature, e.g., different status of smoke exposure). If node i is one of the parents of node j, then the coefficients follow a normal distribution with mean zero and a known large variance (V_c, V_x, or V_y). Otherwise, they have a point mass at zero. Although not the focus of the present work, other prior distributions for $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ can be used as well, for instance, the g-prior (Zellner, 1986; Smith and Kohn, 1996; Fernández et al., 2001; Lee et al., 2003), or the two-component G-prior (Zhang et al., 2016).

Bernoulli $Ber(0.5)$ are chosen for the indicator variables, $\boldsymbol{r}_j^{(c)} = (r_{j1}^{(c)}, r_{j2}^{(c)}, \cdots, r_{j(j-1)}^{(c)})$, $\boldsymbol{r}_j^{(x)} = (r_{j1}^{(x)}, r_{j2}^{(x)}, \cdots, r_{j(j-1)}^{(x)})$, and $\boldsymbol{r}_j^{(y)} = (r_{j1}^{(y)}, r_{j2}^{(y)}, \cdots, r_{j(j-1)}^{(y)})$. With $Ber(0.5)$, we assume no prior knowledge on the inclusion of a parent node and inference on parental nodes selection relies on information in the data. In the situation of available prior knowledge on network structures, instead of 0.5 in $Ber(0.5)$, nodes with low probabilities of being parents can take a value smaller than 0.5. For variance components, we choose vague prior distributions for $\sigma_{x(j)}^2$ and $\sigma_{y(j)}^2$, in particular, an inverse gamma distributions with small shape and scale parameters. As seen in the above definitions, for a given η and ordering $\mathcal{O}$, all the definitions of prior and hyper-prior distributions of the parameters are independent.

So far, the specification of network structures as well as prior distributions on edges such as the regression coefficients are conditional on a given ordering $\mathcal{O}$, and $\mathcal{O}$ needs to be inferred. We assume its prior distribution is uniform among all possible permutations of p nodes, and propose an efficient posterior sampling approach in the next session to infer $\mathcal{O}$.

2.3. The joint posterior distribution and its computing

For a given graph ordering $\mathcal{O}$, to estimate $\boldsymbol{\beta}_{ij}^{(1)}$ and $\boldsymbol{\beta}_{ij}^{(2)}$, a set of prior and hyper-prior parameters need to be inferred, including parameters when $\eta = 0$: $\boldsymbol{\beta}_j^{(x)} = (\beta_{j1}^{(x)}, \beta_{j2}^{(x)}, \cdots, \beta_{j(j-1)}^{(x)}), \boldsymbol{r}_j^{(x)}$, $\boldsymbol{\beta}_j^{(y)} = (\beta_{j1}^{(y)}, \beta_{j2}^{(y)}, \cdots, \beta_{j(j-1)}^{(y)}), \boldsymbol{r}_j^{(y)}$, pa-

rameters when $\eta = 1$: $\boldsymbol{\beta}_j^{(c)} = (\beta_{j1}^{(c)}, \beta_{j2}^{(c)}, \cdots, \beta_{j(j-1)}^{(c)}), \boldsymbol{r}_j^{(c)}$, as well as the variance components $\sigma^2_{x(j)}$ and $\sigma^2_{y(j)}$. Under the context of Bayesian inferences via Markov chain Monte Carlo (MCMC) simulations, we combine these parameters along with the ordering of nodes $\mathcal{O}$ into a collection of parameters that fit into different states of η, $\boldsymbol{\theta} = (\mathcal{O}, \boldsymbol{\beta}_j^{(x)}, \boldsymbol{r}_j^{(x)}, \boldsymbol{\beta}_j^{(y)}, \boldsymbol{r}_j^{(y)}, \eta, \boldsymbol{\beta}_j^{(c)}, \boldsymbol{r}_j^{(c)}, \sigma^2_{x(j)}, \sigma^2_{y(j)}, j = 1, \cdots, p)$. Inference on this collection of parameters will produce an estimate of the networks and assess the probability of having identical networks. The joint likelihood of $\boldsymbol{\theta}$ is

$$
\begin{aligned}
&L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) \\
&\propto \prod_{j=1}^{p} \left\{ (\sigma^2_{x(i)})^{-\frac{n_x}{2}} \exp\left[-\frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)}{2\sigma^2_{x(j)}} \right] \right. \\
&\left. \times (\sigma^2_{y(j)})^{-\frac{n_y}{2}} \exp\left[-\frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i)}{2\sigma^2_{y(j)}} \right] \right\},
\end{aligned}
$$

with $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ defined in (3).

The joint posterior distribution of $\boldsymbol{\theta}$ is,

$$
\begin{aligned}
&p(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) \\
&\propto \prod_{j=1}^{p} \left\{ (\sigma^2_{x(j)})^{-\frac{n_x}{2}} \exp\left[-\frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)}{2\sigma^2_{x(j)}} \right] \right. \\
&\times (\sigma^2_{y(j)})^{-\frac{n_y}{2}} \exp\left[-\frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_j)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i)}{2\sigma^2_{y(j)}} \right] \\
&\times \prod_{i=1}^{j-1} \left[p(\beta_{ij}^{(c)}|\eta = 1, \mathcal{O}, r_{ij}^{(c)}) p(\beta_{ij}^{(x)}|\eta = 0, \mathcal{O}, r_{ij}^{(x)}) \right. \\
&\left. \left. \times p(\beta_{ij}^{(y)}|\eta = 0, \mathcal{O}, r_{ij}^{(y)}) p(r_{ij}^{(c)}|\eta = 1) p(r_{ij}^{(x)}|\eta = 0) p(r_{ij}^{(y)}|\eta = 0) p(\sigma^2_{x(j)}) p(\sigma^2_{y(j)}) \right] \right\} \\
&\times p(\eta) p(\mathcal{O}). \qquad (4)
\end{aligned}
$$

The Gibbs sampler is applied to full conditional posterior distributions of each parameter in $\boldsymbol{\theta}$ to sequentially draw posterior samples, based on which we infer η, graph structure determined by $r_{ij}^{(c)}$ if $\eta = 1$ or $r_{ij}^{(x)}$ and $r_{ij}^{(y)}$ if $\eta = 0$,

and $\boldsymbol{\beta}_{ij}^{(1)}$ and $\boldsymbol{\beta}_{ij}^{(2)}$ describing the strength of connections between nodes. In the following subsections, we present and discuss conditional posterior distributions of the parameters.

2.3.1. Conditional posterior probability of η

Since the decision on whether two networks are differential or not is critical to the estimates of network structure and corresponding parameters, we start from presenting the conditional posterior of η. Denoted by $(\cdot)$ a collection of all conditional parameters, based on (4), we have

$$p(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) \quad \propto \quad p(\boldsymbol{X}, \boldsymbol{Y}|\eta = 1, \boldsymbol{r}^{(c)}, \boldsymbol{\sigma}_x^2, \boldsymbol{\sigma}_y^2, \mathcal{O})p(\eta = 1),$$

where $\boldsymbol{\sigma}_x^2 = \{\sigma_{x(j)}^2, j = 1, \cdots, p\}$ and $\boldsymbol{\sigma}_y^2 = \{\sigma_{y(j)}^2, j = 1, \cdots, p\}$.

It can be shown that, when n_x and n_y large, the full conditional posterior probability, $p(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$, can be approximated by the following (Appendix I),

$$\begin{aligned} p(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) &\approx \Big[1 + \exp\{\log(b_\eta) - \log(a_\eta) + \lambda(n)\}\Big]^{-1}, \\ \lambda(n) &= 1/2(|E| \log n - |E_x| \log n_x - |E_y| \log n_y), \end{aligned}$$

where $n = n_x + n_y$. In the above, $|E|, |E_x|,$ and $|E_y|$ denote numbers of edges in inferred identical and differential networks, respectively, and

$$\begin{aligned} a_\eta &= p(\boldsymbol{X}, \boldsymbol{Y}|(\cdot), \eta = 1) \\ &\propto \prod_{j=1}^{p} \Bigg\{ (\sigma_{x(j)}^2)^{-\frac{n_x}{2}} \exp\Bigg[- \frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{X}_i)}{2\sigma_{x(j)}^2} \Bigg] \\ &\quad \times (\sigma_{y(j)}^2)^{-\frac{n_y}{2}} \exp\Bigg[- \frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)}{2\sigma_{y(j)}^2} \Bigg] \Bigg\}, \quad (5) \\ b_\eta &= p(\boldsymbol{X}, \boldsymbol{Y}|(\cdot), \eta = 0) \\ &\propto \prod_{j=1}^{p} \Bigg\{ (\sigma_{x(j)}^2)^{-\frac{n_x}{2}} \exp\Bigg[- \frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)}{2\sigma_{x(j)}^2} \Bigg] \\ &\quad \times (\sigma_{y(j)}^2)^{-\frac{n_y}{2}} \exp\Bigg[- \frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(y)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(y)} \boldsymbol{Y}_i)}{2\sigma_{y(j)}^2} \Bigg] \Bigg\}. \quad (6) \end{aligned}$$

We denote this approximated conditional posterior probability as $p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$ and it has the following Proposition.

Proposition: Assume 1) sparse networks with $|E|, |E_x|$, and $|E_y|$ in the order of $O(p)$, 2) $n_x \to \infty$ and $n_y \to \infty$ in the same speed, and 3) $\log n_x/p \to \infty$ as $n_x, p \to \infty$, and similar assumptions applied to n_y. Then $\lim_{n_x, n_y \to \infty} p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 1$ if the underlying $\eta = 1$, and $\lim_{n_x, n_y \to \infty} p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 0$ if the underlying $\eta = 0$.

The proof of the Proposition is in Appendix II. This Proposition indicates that, with $p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$, the underlying truth of η will be identified asymptotically. In addition, for network constructions, $p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$ has a potential to penalize large numbers of edges as indicated by the definition of $\lambda(n)$. In genetic and epigenetic studies, this property benefits marker detection and is practically informative to clinicians and health researchers. In the context of network comparisons, the definition of $\lambda(n)$ in $p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$ implies a preference for identical networks over differential networks. The Proposition holds for other choices of prior distributions of the parameters as long as the conditional priors of $\beta_{ij}^{(1)}$ and $\beta_{ij}^{(2)}$ are non-informative for parental node i, i.e., nodes such that $r_{ij}^{(\cdot)} = 1$. Although not the situation in our proposed method as seen from the Proposition of $p^{\lambda}(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$, the Jeffreys−Lindley paradox suggests that a caution should be made in any hypothesis testing conducted under the Bayesian framework, since non-informative prior distributions can possibly lead to rather strong but useless decision, e.g., rejection of null with probability 1 regardless of data (Robert, 2007).

2.3.2. Conditional posterior distributions of other parameters

Below, we list the conditional posterior distributions for the remaining parameters. For the parameters to select a parent node k at a child node j, when $\eta = 1$,

$$p(r_{jk}^{(c)} = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}, \eta = 1) = \frac{a_c}{a_c + b_c},$$

where a_c and b_c are proportional to the conditional posterior distributions of $r_{jk}^{(c)} = 1$ and $r_{jk}^{(c)} = 0$, respectively,

$$a_c = \exp\left\{ -\frac{(\beta_{jk}^{(c)})^2}{2V_c} - \frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{X}_i)}{2\sigma_{x(j)}^2} - \frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)}{2\sigma_{y(j)}^2} \right\} \times p(r_{jk}^{(c)} = 1),$$

and

$$b_c = \exp\left\{ -\frac{(\boldsymbol{X}_j - \sum_{i=1, j\neq k}^{i-1} \beta_{ij}^{(c)} \boldsymbol{X}_j)^T (\boldsymbol{X}_i - \sum_{j=1, j\neq k}^{i-1} \beta_{ij}^{(c)} \boldsymbol{X}_j)}{2\sigma_{x(i)}^2} - \frac{(\boldsymbol{Y}_j - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)}{2\sigma_{y(j)}^2} \right\} \times p(r_{jk}^{(c)} = 0).$$

When $\eta = 0$, each population holds its own network and the conditional posterior distribution of $r_{jk}^{(x)}$ is defined as

$$p(r_{jk}^{(x)} = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}, \eta = 0) = \frac{a_x}{a_x + b_x},$$

with $a_x = \exp\left\{ -\frac{\beta_{jk}^{(x)}}{2V_x} - \frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)}{2\sigma_{x(j)}^2} \right\} \times p(r_{jk}^{(x)} = 1)$, and $b_x = \exp\left\{ -\frac{(\boldsymbol{X}_j - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)^T (\boldsymbol{X}_i - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(x)} \boldsymbol{X}_i)}{2\sigma_{x(j)}^2} \right\} \times p(r_{jk}^{(x)} = 0)$. The conditional posterior of $r_{jk}^{(y)}$ is in a similar form.

Turning to the regression coefficients, if $r_{jk}^{(\cdot)} = 0$, then $\beta_{jk}^{(\cdot)} = 0$. Otherwise, the conditional posterior distribution of $\beta_{jk}^{(c)}$ is univariate normal, $N(\mu_{\beta_{jk}^{(c)}}, \sigma_{\beta_{jk}^{(c)}}^2)$, with $\mu_{\beta_{jk}^{(c)}} = (V_c \sigma_{y(j)}^2 \boldsymbol{C}_1^T \boldsymbol{X}_k + V_c \sigma_{x(j)}^2 \boldsymbol{C}_2^T \boldsymbol{Y}_k)/(\sigma_{x(j)}^2 \sigma_{y(j)}^2 + V_c \sigma_{y(j)}^2 \boldsymbol{X}_k^T \boldsymbol{X}_k + V_c \sigma_{x(j)}^2 \boldsymbol{Y}_k^T \boldsymbol{Y}_k)$, $\sigma_{\beta_{jk}^{(c)}}^2 = (\sigma_{x(j)}^2 \sigma_{y(j)}^2 V_c)/(\sigma_{x(j)}^2 \sigma_{y(j)}^2 + V_c \sigma_{y(j)}^2 \boldsymbol{X}_k^T \boldsymbol{X}_k + V_c \sigma_{x(j)}^2 \boldsymbol{Y}_k^T \boldsymbol{Y}_k)$, where $C_1 = (\boldsymbol{X}_j - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(c)} \boldsymbol{X}_i)$ and $C_2 = (\boldsymbol{Y}_j - \sum_{i=1, i\neq k}^{j-1} \beta_{ij}^{(c)} \boldsymbol{Y}_i)$. The conditional posterior distributions of $\beta_{jk}^{(x)}|(\cdot), \boldsymbol{X}, \boldsymbol{Y}$ and $\beta_{jk}^{(y)}|(\cdot), \boldsymbol{X}, \boldsymbol{Y}$ are defined in a similar way.

Finally, we discuss the conditional posterior distribution of $\mathcal{O}$, ordering of the nodes, and its sampling. From the joint posterior distribution of $\boldsymbol{\theta}$ given in (4), derivation of the conditional posterior distribution of $\mathcal{O}$ is straightforward, which is,

$$
\begin{aligned}
&p(\mathcal{O}|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) \\
&\propto \prod_{j=1}^{p} \Bigg\{ (\sigma^2_{x(j)})^{-\frac{n_x}{2}} \exp\left[- \frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)^T (\boldsymbol{X}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(1)} \boldsymbol{X}_i)}{2\sigma^2_{x(j)}} \right] \\
&\times (\sigma^2_{y(j)})^{-\frac{n_y}{2}} \exp\left[- \frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i)^T (\boldsymbol{Y}_j - \sum_{i=1}^{j-1} \beta_{ij}^{(2)} \boldsymbol{Y}_i)}{2\sigma^2_{y(j)}} \right] \\
&\times \prod_{i=1}^{j-1} \left[p(\beta_{ij}^{(1)}|\eta, \mathcal{O}, r_{ij}^{(\cdot)}) p(\beta_{ij}^{(2)}|\eta, \mathcal{O}, r_{ij}^{(\cdot)}) \right] \Bigg\},
\end{aligned}
$$

where $r_{ij}^{(\cdot)}$ is $r_{ij}^{(x)}$, $r_{ij}^{(y)}$, or $r_{ij}^{(c)}$.

2.3.3. Sampling of graph ordering

Since the number of nodes p can be large, an efficient sampling of $\mathcal{O}$ that has the ability to escape from traps of local maximum is critical in practice. Energy-driven sampling has been used often to diminish this type of concern (Ellis and Wong, 2008; Van den Bergh et al., 2012). We adopt the sampling scheme suggested in Han et al. (2016), the Adjusted Single Queue Equi-Energy algorithm (ASQEE), which is adapted from the SQEE sampling method proposed by Ellis and Wong (2008).

Basically, the SQEE approach utilizes energy and energy rings with minimum energy suggested by the range of $H(\mathcal{O}) = -\log(p(\mathcal{O}|(\cdot), \boldsymbol{X}, \boldsymbol{Y}))$, allowing energy upper bound to be ∞, and energy rings formed by dividing the range of energy into groups (or "chains" as in Ellis and Wong (2008)). Energy levels increase from lower to upper rings, and within each ring, the probability density function is $\pi_l(\mathcal{O}) = \exp(\frac{-\max\{H(\mathcal{O}), H_l\}}{T_l}), l = 1, 2, \cdots, W$, with l indexing groups or chains and in total W groups (thus W rings), H_l is lower bound energy level for chain l, and T_l is the lower temperature of that chain such that

$1 = T_1 < T_2 < \cdots < T_W$. A ring in group l, D_l, is a collection of different orderings such that their energy is bounded by corresponding lower and upper bound energy levels, $D_l = \{\mathcal{O}|H(\mathcal{O}) \in [H_l, H_{l+1}), l = 1, \cdots, W\}$ with H_{W+1} set at ∞. In our study (both simulations and real data applications), we take $W = 5$ to allow Markov Chain Monte Carlo (MCMC) sampling between rings for the purpose of fast convergence to the global maximum. This construction shows that when $l = 1$, $\pi_1(\mathcal{O})$ is the target distribution. Furthermore, as the value of l increases, the distribution in the lth group is more flatten, enhancing the ability of the chain jumping across different modes to avoid being trapped at local maximums. To perform the sampling, we follow the "cylindrical shift" operation suggested in Ellis and Wong (2008) to propose an ordering. Then a Metropolis-Hastings (Hastings, 1970) algorithms is applied to determine whether the newly proposed ordering will be accepted or not, which is the standard local Metropolis-Hastings move. The sampling starts from the chain with the highest energy level, which is associated with a flatten distribution. This allows the chains to move more quickly through the space to collect samples for later communications with lower-temperature chains.

For the sampling scheme ASQEE in Han et al. (2016), when evaluating the conditional probability of a sampled ordering, instead of utilizing all possible graphs for that order, it estimated the probability based on a graph showing the highest probability for a given graph ordering aiming to improve sampling efficiency. Readers are referred to Ellis and Wong (2008) and Han et al. (2016) for detailed discussions on the ASQEE sampler construction and its related algorithms.

3. Numerical Analysis

Via simulations, we demonstrate finite sample properties of the proposed method under different scenarios and compare the findings with those from existing methods that can be applied to test network differentiation.

3.1. Simulation scenarios

Generating Monte Carlo (MC) replicates: We consider DNA methylation measures, $\boldsymbol{X}$ and $\boldsymbol{Y}$, from two populations (e.g., exposed vs. non-exposed to in utero smoking) and each with $n_x = n_y$ observations. Each data set is generated from an underlying network structure with p CpG sites (or p nodes) and $|E_{0x}|$ and $|E_{0y}|$ edges, respectively, based on linear regressions. We assume each node can have up to four parents corresponding to regression coefficients of $\boldsymbol{\beta} = \{1.5, 2, 2, 2.5\}$ in order. The root node is an experimental node and does not have any parents. Two types of underlying networks are considered. The first is that the two populations share the same networks (i.e., identical networks) and the other situation is that each population has its unique network (i.e., differential networks). In our simulations, we take $p = 10, 20$ and $n_x = n_y = 50, 100, 200$. For identical networks, we choose $|E_{0x}| = |E_{0y}| = |E_{0c}| = 10, 20$, and for differential networks, we consider two sets of $|E_{0x}|$ and $|E_{0y}|$, $|E_{0x}| = 5, |E_{0y}| = 10$ and $|E_{0x}| = 20, |E_{0y}| = 10$. For each graph, a level of sparsity is defined as the ratio between the true number of edges and the possible number of edges, $S = 2|E_{\cdot}|/(p(p-1))$, where $|E_{\cdot}|$ represents $|E_{0x}|, |E_{0y}|$, or $|E_{0c}|$. For instance, a graph with 10 nodes and 10 edges has a level of sparsity $10/45 = 0.222$. The connection of each edge is randomly selected based on a prespecified ordering of all the nodes. The random error when generating each node is assumed to be normally distributed with mean 0 and variance 1. For each combination of the settings of $n_x = n_y$, p, E_{0x} and E_{0y}, we generated 100 MC replicates.

Posterior sampling: For each MC replicate, we utilize the Gibbs sampler to sequentially draw samples of each parameter from its conditional posterior probability density (or mass) function. Since one ordering can lead to a number of graphs, when estimating conditional posterior probability of a sampled ordering, we run a set of iterations aiming to capture graphs with high probabilities for that given ordering. In addition, to increase the stability of sampled ordering, in each energy ring, we sample a series of orderings as burn-in following the SQEE and ASQEE sampling technique. To get an insight on the numbers of iterations needed for these considerations, we first run longer chains using the

Gibbs sampler on several MC replicates and examine the quality of posterior inferences. After observing fast convergence with respect to the inference of η and graph structure, for each of the 100 MC replicates, we run 6,000 iterations which includes 50 iterations for inferring conditional posterior probabilities of each sampled ordering with 25 as burn-in iterations and 120 iterations for sampling orderings with a range of 20 to 100 iterations (higher energy rings with less iterations) as burn-in iterations across 5 energy rings.

Summarizing statistics: Graphs and orderings of nodes are not one-to-one correspondence and one graph can be a result of multiple orderings. Since our goal is to compare graphs between two populations, our posterior inferences focus on graphs rather than ordering of graphs. Four statistics focusing on testing and network constructions are used to summarize the results and assess the proposed method: 1) power of correct detection with respect to network comparison (identical or differential), 2) average proportion of true positives for edge connection and directions (TPCD) in a network, 3) average proportion of false positives (FP) of a network, and 4) average proportion of correct connections (CC) of edges. A proportion of correct connections combines information on sensitivity (reflected by proportions of true positives) and specificity (reflected by proportions of false positives). For all the statistics except for power, we also infer 95% empirical intervals. We evaluate the proposed method based on these statistics on the various choices of sample sizes and numbers of edges noted in the paragraph above.

Competing methods: Approaches that not only compare networks but also infer networks are relatively limited. To assess the proposed method, we use two existing approaches, one focuses on comparisons in structures between two networks and the other on coefficients comparisons. The first competing method is proposed by Almudevar (2010). It compares two graphs with each constructed based on minimum spanning trees (MST) and utilizes permutations to calculate an empirical p-value for decision-making. We denote this method as MST-based approach. In the second comparison, we utilize an existing method in multivariate testing, the Hotelling's T-squared test. In particular, we first infer networks

for the two populations separately using the network construction method implemented in the proposed approach, and then apply the Hotelling's T-squared test on the posterior samples of regression coefficients assuming unequal variances between the two populations. Posterior samples are selected to ensure small values on autocorrelation functions. In both comparisons, we compare the power of detecting underlying truth using each of the competing approach with that from the proposed method.

3.2. Results

Table 1 lists different model assessment statistics when in total 10 nodes are considered. In the situation that the underlying networks are identical ($\eta = 1$), overall the power of detecting the correct type of networks (identical or differential) is reasonably high for all cases. Since the underlying networks are identical, the decrease in power when the sample size is large shown in the table is a phenomenon observed in a two sample hypothesis testing when the null (i.e., two means are equal) is true. This consequently caused other assessment statistics being slightly inferior. We note that the false positives are influenced by the edge sparsity of the networks. When the graphs are sparse (e.g., 10 nodes with 10 edges with sparsity 0.222), proportions of false positives are low but larger false positives are observed when the graphs are less sparse (e.g., 10 nodes with 20 edges with sparsity 0.444). All these are likely due to the inclusion of edges that are indirectly connected to a node under investigation, for instance, by being a "parent" of this node's "child", a phenomenon discussed in Wasserman and Roeder (2009). All these lead to an overall slight decrease in the proportions of correctness (top left panel of Figure 1) when the number of edges is large and sparsity is low, and this type of patterns continues when the number of nodes is 20 (top right panel of Figure 1).

On the other hand, when underlying networks are two differential networks ($\eta = 0$), overall the power to detect the truth is higher than the power when the underlying networks are identical. In addition, proportions of TPCD increase with the increase of sample size. As in the situation of $\eta = 1$, false positives

slightly increase as sparsity level decreases (Table 1 and bottom panel of Figure 1), leading to decrease in proportions of correctness.

Since the concept of correctness combines both sensitivity and specificity, we examine this statistics a little further. As reflected by the patterns shown in Figure 1, with the number of nodes and sample size fixed, sparsity seems to play an important role in the determination of proportion of correctness, regardless of the number of edges; the lower the sparsity (i.e., high sparsity values), the lower the proportion of correctness. On the other hand, smaller numbers of nodes lead to higher proportions of correctness for similar sparsity levels (demonstrated by results with sparsity of 0.11 shown in the two figures at the lower panel of Figure 1).

To further evaluate the approach, we next compare the results from the proposed method with those from the two competing approaches, the MST-based approach and the approach based on Hotelling's T-squared tests. Since the MST-based approach is designed for small sample sizes, we used the MC replicates under the setting of $p = 10$ nodes and each MC replicate having $n_x = n_y = 50$ observations. When underlying two networks are identical, the power of detecting this underlying truth is 0.99. However, the proportions of true positives, false positives, and correct connections, along with 95% empirical intervals, are $0.40(0.10, 0.60)$, $0.14(0.086, 0.23)$, and $0.76(0.62, 0.84)$, respectively, all inferior to the corresponding results in Table 1 (first row in the first block of Table 1). When underlying two networks are differential, the power is only 37%, substantially lower than the power from the proposed approach (first row in the third block of Table 1).

For the second competing method based on Hotelling's T-squared tests, we present the results of power assessment using the MC replicates generated under the settings with $p = 20$. Overall, when underlying two networks are identical, the power to detect the truth is much lower than that from the proposed approach, although the pattern is the same, i.e., the power decreases with the increase of sample sizes (Figure 2). As expected, when two networks are truly differential, the power of detecting the truth increases with sample sizes and is

Table 1: Summary statistics for detecting differential networks, including estimated power of correct detection (with respect to network types), true positives for edge connections and directions (TPCD), false positives (FP), and correct connections (CC) across 100 MC replicates along with 95% empirical intervals (EI).

Sample size ($n_x = n_y$)	Power (%)	TPCD (95% EI)	FP (95% EI)	CC (95% EI)
		Underlying networks: identical networks ($p = 10$ nodes, $\|E_{0c}\| = 10$ edges)		
50	94.9	0.999 (0.998, 1.0)	0.020 (0.0, 0.086)	0.984 (0.933, 1.0)
100	98.2	0.999 (0.999, 1.0)	0.021 (0.0, 0.086)	0.984 (0.932, 1.0)
200	98.9	0.999 (0.998, 1.0)	0.013 (0.0, 0.072)	0.990 (0.944, 1.0)
		Underlying networks: identical networks ($p = 10$ nodes, $\|E_{0c}\| = 20$ edges)		
50	90.0	0.984 (0.900, 1.0)	0.199 (0.0, 0.622)	0.882 (0.644, 1.0)
100	91.7	0.996 (0.950, 1.0)	0.199 (0.0, 0.560)	0.887 (0.676, 1.0)
200	88.7	0.993 (0.965, 1.0)	0.230 (0.0, 0.640)	0.869 (0.629, 1.0)
		Underlying networks: differential networks ($p = 10$ nodes, $\|E_{0x}\| = 5, \|E_{0y}\| = 10$ edges)		
50	99.9	X : 0.998 (0.999, 1.0)	0.017 (0.00, 0.075)	0.984 (0.933, 1.0)
		Y : 0.977 (0.80, 1.0)	0.047 (0.00, 0.200)	0.958 (0.844, 1.0)
100	99.9	X : 1.0 (0.999, 1.0)	0.016 (0.00, 0.088)	0.986 (0.921, 1.0)
		Y : 0.992 (0.90, 1.0)	0.045 (0.00, 0.329)	0.963 (0.744, 1.0)
200	99.9	X : 1.0 (0.999, 1.0)	0.013 (0.00, 0.063)	0.988 (0.944, 1.0)
		Y : 0.996 (0.998, 1.0)	0.048 (0.00, 0.287)	0.962 (0.776, 1.00)
		Underlying networks: differential networks ($p = 10$ nodes, $\|E_{0x}\| = 20, \|E_{0y}\| = 10$ edges)		
50	99.9	X : 0.979 (0.874, 1.0)	0.218 (0.0, 0.560)	0.869 (0.667, 1.0)
		Y : 0.976 (0.90, 1.0)	0.039 (0.0, 0.171)	0.964 (0.867, 1.0)
100	99.9	X : 0.993 (0.950, 1.0)	0.213 (0.0, 0.600)	0.879 (0.633, 1.0)
		Y : 0.991 (0.90, 1.0)	0.033 (0.0, 0.186)	0.972 (0.855, 1.0)
200	99.9	X : 0.993 (0.950, 1.0)	0.232 (0.0, 0.600)	0.868 (0.644, 1.0)
		Y : 0.998 (0.997, 1.0)	0.040 (0.0, 0.230)	0.967 (0.821, 1.0)

overall high but lower than the power based on the proposed method. The findings with $p = 10$ follow the same trend but were deteriorate when two networks are truly identical. The proposed approach clearly outperforms the method built upon the Hotelling's T-squared test.

4. Real Data Application

We apply the method to DNA methylation of 23 CpG sites in 9 genes (Table 2) analyzed in our epigenetic epidemiological study. Each of these CpGs was

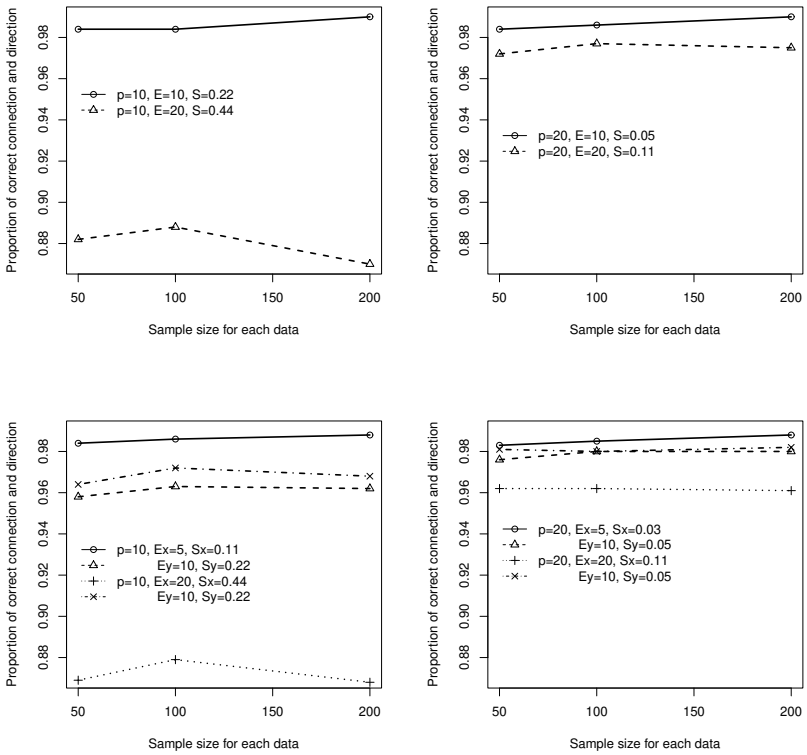


Figure 1: Plots of proportions of correct connections with correct direction of connections. The top panel is for identical networks and the lower panel is for differential networks.

shown to be associated with maternal smoking during pregnancy (Joubert et al., 2012). DNA methylation data of 245 girls measured at age 18 are used in the analysis. These 245 subjects represent a random sample from the Isle of Wight birth cohort (Arshad et al., 2018; Quraishi et al., 2015). Among these 245 girls, 48 were exposed to maternal smoking during pregnancy. We demonstrate the proposed method from two aspects. Firstly, we only consider the 197 subjects not exposed to maternal smoking during pregnancy. We disturb the data by introducing noise to the first 97 subjects on one CpG site (cg18146737 [node 15] on gene *GFI1*) to artificially produce two conditions, one for the first 97 subjects and the other for the remaining 100 subjects. This disturbance is expected

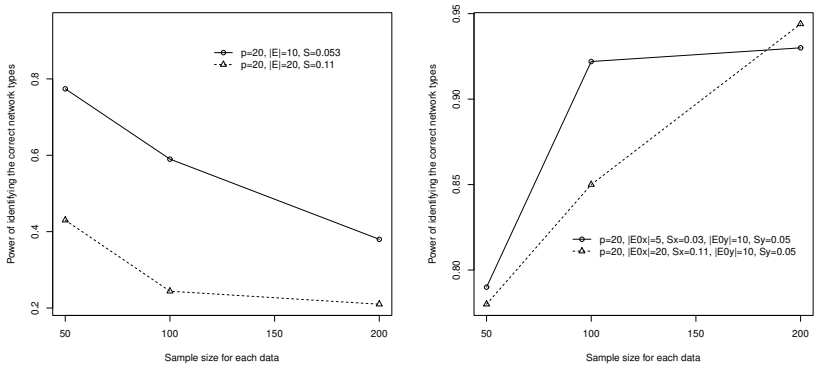


Figure 2: Power of detecting the underlying truth of network differentiation status using a method based on the Hotelling's T-squared tests. Left panel: two networks are truly identical. Right panel: two networks are truly differential.

to break the links connected to the CpG site cg18146737, which should lead to underlying two differential networks among the CpGs. We then apply the developed method to test whether the two networks are identical or differential. In the second aspect, to demonstrate our approach in the real world, we apply the method to all 245 subjects and assess whether the network among the CpGs for the subjects whose mother did not smoke during pregnancy is differential compared to the network for the subjects whose mother smoked. For each scenario, we run 84,000 MCMC iterations, which includes 200 iterations used to calculate conditional posterior probabilities of each sampled ordering with 100 as burn-in iterations and 4,200 iterations for sampling ordering with a range of 70 to 350 iterations as burn-in iterations across 5 energy rings.

In the first scenario with disturbance given to the first 97 subjects, the inferred posterior probability that the two networks are differential is 0.95, implying a high potential that the two networks are differential. However, there is a possibility that the first 97 subjects were under an unknown condition different from the remaining 100 subjects, and thus the underlying networks were already differential even before we disturb the data. To test this, we use the original data for the 197 subjects without disturbing the data but still assume

Table 2: The list of CpGs and their corresponding genes

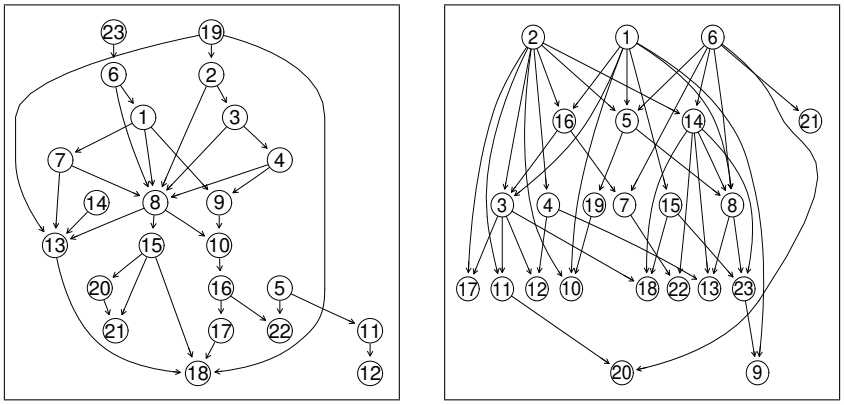
Label	CpG	Gene	Label	CpG	Gene
1	cg03991871	*AHRR*	13	cg14179389	*GFI1*
2	cg04180046	*MYO1G*	14	cg18092474	*CYP1A1*
3	cg04598670	*ENSG00000225718*	15	cg18146737	*GFI1*
4	cg05549655	*CYP1A1*	16	cg18316974	*GFI1*
5	cg05575921	*AHRR*	17	cg18655025	*TTC7B*
6	cg06338710	*GFI1*	18	cg19089201	*MYO1G*
7	cg10399789	*GFI1*	19	cg21161138	*AHRR*
8	cg11715943	*HLA-DPB2*	20	cg22132788	*MYO1G*
9	cg11924019	*CYP1A1*	21	cg22549041	*CYP1A1*
10	cg12477880	*RUNX1*	22	cg23067299	*AHRR*
11	cg12803068	*MYO1G*	23	cg25949550	*CNTNAP2*
12	cg12876356	*GFI1*			

two conditions between the first 97 subjects and the remaining 100 subjects. After applying the method to the original data without disturbance, the posterior probability of having identical networks is 0.66, indicating that the 197 subjects are likely sharing the same network.

In the second scenario, we apply the method to all the 245 subjects. The posterior probability of having differential networks is 0.99, suggesting that subjects exposed to maternal smoking during pregnancy and subjects not exposed are highly likely to have their unique networks. The inferred networks for both groups are shown in Figure 3.

Comparing the two networks (smoke exposed vs. smoke non-exposed) inferred based on data of 245 subjects, we observed substantially reduced connections of nodes 1 (cg03991871 on *AHRR*), 2 (cg04180046 on *MYO1G*), and 6 (cg06338710 on *GFI1*) as root nodes in the network for subjects with in utero smoke exposure. These CpGs are potential driving factors important to the differentiation between the two networks, and deserve further laboratory ex-

aminations and investigations. Another node 14 with cg18092474 (*CYP1A1*) also draws our attention. Although it is not like nodes 1, 2, and 6 such that none of these three nodes have parent nodes, node 14 also has a large number of connections with its children in the network for non-exposed subjects but only one child in the other network. In a recent meta analyses (Joubert et al., 2016), DNA methylation at these CpGs were demonstrated to be strong markers for in utero smoke exposure. In another study, cg03991871, cg04180046, and cg18092474, along with other CpG sites, are used to predict status of smoke exposure and the accuracy is 81% (Ladd-Acosta et al., 2016). To our knowledge, the inter-connections between these genes and DNA methylation sites have not been examined in any other studies. The findings from this real data application provide a potential and necessity for future investigations on the potential regulatory functionality of these four CpGs and the genes to which they are mapped. Additionally, instead of examining all possible CpG sites related to maternal smoke exposure during pregnancy, using the proposed method to assess differentiation and to select CpGs potentially leading to differentiation will substantially reduce the laboratory burden and make the experiment easier to manage.



a) b)

Figure 3: Estimated two differential networks . a) Subjects in utero exposed to smoke (48 subjects) b) Subjects not exposed (197 subjects)

5. Summary and Discussion

To examine the differentiation of joint activities for a set of CpG sites between two groups (exposed vs. non-exposed to smoke in utero) and identify potential driving factors leading to the differentiation, we utilized Bayesian networks and proposed a Bayesian method built upon the concept of variable selection to conclude the status of differentiation. The approximated conditional posterior probability mass function for the decision indicator variable has the property to converge to the underlying truth in terms of network differentiation. In the process of testing network differentiation, we estimated graph ordering using the Adjusted Single Queue Equi-Energy (ASQEE) algorithm proposed by Han et al. (2016) for the purpose to escape from local maximums. Theoretical assessment and simulations have demonstrated the effectiveness of the proposed methods in assessing differential networks. Real data applications further demonstrated that the method is practically useful and effective. We identified four potentially driving epigenetic factors, cg03991871, cg04180046, cg06338710 and cg18092474, such that they have the largest numbers of children.

Note that both $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the proposed approach represent observational data and no experimental data are assumed. In this case, parents of node i are inferred based on posterior probability of edge connection indicators (i.e., $r_{ik}^{(c)}$, $r_{ik}^{(x)}$, and $r_{ik}^{(y)}$) conditional on other parameters. Without experimental data, one needs to be aware that only Markov equivalent networks are constructed for each given ordering of nodes (Andersson et al., 1997). Although this will not affect our conclusion on network differentiation (since given an ordering the true graph is among all the Markov equivalent networks for that ordering), the inferred network may not be the underlying true network. In practice, one way to ease this uncertainty is to bring in expertise from the corresponding research field, e.g., biologist and epigenetic epidemiologist for the study in our motivating example, to choose a network that is most practically meaningful.

The proposed approach for network comparisons utilizes an indicator variable for status of network differentiation, and has the ability to estimate net-

works as well as compare networks. If the focus is only on network comparisons, then we may consider augmenting the data by adding a node denoting treatment status, in which case no treatment effects indicate that the two networks are identical. This type of data augmentation, however, can be challenging for researchers in the applied fields, e.g., geneticists, since this added node is not a stimulus or experimental node but to assist statistical modeling. In addition, through this data augmentation, the strength and direction of connections may have to be estimated separately, should two networks conclude to be differential. In order to infer network and do comparison together, one way is to include interaction effects of treatment with each candidate parent. A careful design and efficient computing algorithms, however, are desired, which surely deserves further investigations.

The new method is not limited to DNA methylation data and is ready to other types of data with ordering unknown, e.g., expression of genes. For ordered data, the method can be easily simplified to fit the situation. In addition, it can be directly applied to perform pair-wise comparisons between multiple networks (> 2 networks), in which case adjustment of multiple testing needs to be considered. An analysis-of-variance-type network testing is desired for differentiation of more than two networks, although this extension maybe computationally intensive. Durante et al. (2018) proposed a Bayesian approach to test the association of undirected networks with a feature of interest, e.g., the association of brain connectivity structures with creative reasoning. This type of hypothesis testing has the potential to fit the needs of comparing more than two directed networks.

Acknowledgements

The research work of H Zhang, W Karmaus, H Arshad, and J Holloway was supported by NIH/NIAID R01AI121226 (MPI: Zhang, Holloway). The work of FI Rezwan was supported by the Ageing Lungs in European Cohorts (ALEC) Study (EU Horizon 2020, Grant number 633212). The authors are thankful to

the High Performance Computing facility at the University of Memphis.

References

Appendix I. The conditional posterior probability of $\boldsymbol{\eta}$, $\boldsymbol{p(\eta = 1|(\cdot), X, Y)}$

In the following, we provide a derivation of the approximated full conditional posterior of η,

$$\begin{aligned} p(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) &\approx \Big[1 + \exp\{\log(\mathrm{b}_\eta) - \log(\mathrm{a}_\eta) + \lambda(n)\}\Big]^{-1} \\ \lambda(n) &= 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y). \end{aligned}$$

It is an approximation of the posterior distribution of η conditional on $\boldsymbol{r}^{(c)}$ and variance components, $\boldsymbol{\sigma}_x^2$ and $\boldsymbol{\sigma}_y^2$, for a given graph ordering $\mathcal{O}$ with $\boldsymbol{r}^{(c)}$ being a collection of indicators representing inclusion or exclusion of parental nodes at each node. The justification laid out in this session follows in spirit the justification of Bayesian Information Criterion in Neath and Cavanaugh (2012). The conditional posterior probability of $\eta = 1$ conditional on $\boldsymbol{r}^{(c)}$ and the variance components for a given ordering is

$$\begin{aligned} p(\eta = 1|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{r}^{(c)}, \boldsymbol{\sigma}_x^2, \boldsymbol{\sigma}_y^2, \mathcal{O}) &\propto p(\boldsymbol{X}, \boldsymbol{Y}|\eta = 1, \boldsymbol{r}^{(c)}, \boldsymbol{\sigma}_x^2, \boldsymbol{\sigma}_y^2, \mathcal{O})p(\eta = 1) \\ &\propto p(\boldsymbol{X}, \boldsymbol{Y}|\eta = 1, \boldsymbol{r}^{(c)}, \boldsymbol{\sigma}_x^2, \boldsymbol{\sigma}_y^2, \mathcal{O}). \end{aligned}$$

The last proportionality is due to the choice of a non-informative prior in our work, $p(\eta = 1) = 0.5$.

In the following, we omit the dependence on $\mathcal{O}$ and the variance components for notation simplicity, but it needs to be clear that all the derivations are conditional on $\mathcal{O}, \boldsymbol{\sigma}_x^2,$ and $\boldsymbol{\sigma}_y^2$. The distribution of $\boldsymbol{X}, \boldsymbol{Y}$ conditional on $\eta = 1$ and $\boldsymbol{r}^{(c)}$ is,

$$
\begin{aligned}
p(\boldsymbol{X},\boldsymbol{Y}|\eta=1,\boldsymbol{r}^{(c)}) &= p(\boldsymbol{Z}|\eta=1,\boldsymbol{r}^{(c)}) \\
&= \prod_{j=1}^{p} p(\boldsymbol{Z_j}|\eta=1,\boldsymbol{r}_j^{(c)}) \\
&= \prod_{j=1}^{p} \int_{\boldsymbol{\beta}_j^{(c)}} p(\boldsymbol{Z_j}|\boldsymbol{\beta}_j^{(c)},\eta=1,\boldsymbol{r}_j^{(c)}) p(\boldsymbol{\beta}_j^{(c)}|\eta=1,\boldsymbol{r}_j^{(c)}) d\boldsymbol{\beta}_j^{(c)} \\
&= \prod_{j=1}^{p} \int_{\boldsymbol{\beta}_j^{(c)}} L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}) p(\boldsymbol{\beta}_j^{(c)}|\eta=1,\boldsymbol{r}_j^{(c)}) d\boldsymbol{\beta}_j^{(c)},
\end{aligned}
$$

where $\boldsymbol{Z}_j = (\boldsymbol{X}_j^T, \boldsymbol{Y}_j^T)^T, \boldsymbol{\beta}_j^{(c)} = \{\boldsymbol{\beta}_{ij}^{(c)}, i = 1, \cdots, j-1\}$ denotes regression coefficients of connected edges at node i under $\eta = 1$, and $\boldsymbol{\beta}_j^{(c)} \sim N(0, V_c\boldsymbol{I})$ with $\boldsymbol{I}$ an identity matrix and V_c known and large to formulate a non-informative but proper prior distribution for $\boldsymbol{\beta}_j^{(c)}$.

Take the natural logarithm transformation of the likelihood function $L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})$ and perform Taylor expansion at $\hat{\boldsymbol{\beta}_j^{(c)}}$, a consistent estimator of $\boldsymbol{\beta}_j^{(c)}$ such that $\lim_{n\to\infty} \left.\frac{\partial \log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})}{\partial \boldsymbol{\beta}_j^{(c)}}\right|_{\boldsymbol{\beta}_j^{(c)}=\hat{\boldsymbol{\beta}_j^{(c)}}} = 0$. We have, for a large n,

$$
\begin{aligned}
\log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}) &\approx \log L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}) \\
&\quad + \left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right)^T \left.\frac{\partial \log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})}{\partial \boldsymbol{\beta}_j^{(c)}}\right|_{\boldsymbol{\beta}_j^{(c)}=\hat{\boldsymbol{\beta}_j^{(c)}}} \\
&\quad + 1/2\left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right)^T \left.\frac{\partial^2 \log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})}{\partial \boldsymbol{\beta}_j^{(c)} \partial \boldsymbol{\beta}_j^{(c)^T}}\right|_{\boldsymbol{\beta}_j^{(c)}=\hat{\boldsymbol{\beta}_j^{(c)}}} \\
&\quad \times \left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right) \\
&\asymp \log L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}) \\
&\quad + 1/2\left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right)^T \left.\frac{\partial^2 \log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})}{\partial \boldsymbol{\beta}_j^{(c)} \partial \boldsymbol{\beta}_j^{(c)^T}}\right|_{\boldsymbol{\beta}_j^{(c)}=\hat{\boldsymbol{\beta}_j^{(c)}}} \\
&\quad \times \left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right) \\
&= \log L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}) \\
&\quad - 1/2\left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right)^T \left[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\right] \left(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\right),
\end{aligned}
$$

where "$\asymp$" denotes "asymptotically equal to", and

$$\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)}) = -1/n(\partial^2/\partial\boldsymbol{\beta}_j^{(c)}\partial\boldsymbol{\beta}_j^{(c)^T})\log L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)})\Big|_{\boldsymbol{\beta}_j^{(c)}=\hat{\boldsymbol{\beta}_j^{(c)}}}$$

is the sample Fisher information matrix.

590 Exponentiate both sides,

$$\begin{aligned} L_j(\boldsymbol{\beta}_j^{(c)}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)}) \quad &\approx \quad L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)}) \\ &\quad \times \exp\Big\{-1/2(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}})^T\Big[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)})\Big](\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}})\Big\}, \end{aligned}$$

which gives

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{Y}|\eta = 1, \boldsymbol{r}_j^{(c)}) \quad &= \quad p(\boldsymbol{Z}|\eta = 1, \boldsymbol{r}^{(c)}) \\ &\approx \quad \prod_{j=1}^{p}\int_{\boldsymbol{\beta}_j^{(c)}} L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)}) \\ &\quad \times \exp\Big\{-1/2\Big(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\Big)^T\Big[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)})\Big] \\ &\quad \Big(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\Big)\Big\}p(\boldsymbol{\beta}_j^{(c)}|\eta = 1, \boldsymbol{r}_j^{(c)})d\boldsymbol{\beta}_j^{(c)} \\ &= \quad \prod_{j=1}^{p} L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)}) \\ &\quad \times \int_{\boldsymbol{\beta}_j^{(c)}} \exp\Big\{-1/2\Big(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\Big)^T\Big[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j}, \eta = 1, \boldsymbol{r}_j^{(c)})\Big] \\ &\quad \Big(\boldsymbol{\beta}_j^{(c)} - \hat{\boldsymbol{\beta}_j^{(c)}}\Big)\Big\}p(\boldsymbol{\beta}_j^{(c)}|\eta = 1, \boldsymbol{r}_j^{(c)})d\boldsymbol{\beta}_j^{(c)}. \end{aligned} \tag{7}$$

For the integration in (7),

$$
\begin{aligned}
&\int_{\boldsymbol{\beta}_j^{(c)}} \exp\Big\{-1/2(\boldsymbol{\beta}_j^{(c)}-\hat{\boldsymbol{\beta}_j^{(c)}})^T\Big[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big](\boldsymbol{\beta}_j^{(c)}-\hat{\boldsymbol{\beta}_j^{(c)}})\Big\}p(\boldsymbol{\beta}_j^{(c)}|\eta=1,\boldsymbol{r}_j^{(c)})d\boldsymbol{\beta}_j^{(c)}\\
&= C_0\int_{\boldsymbol{\beta}_j^{(c)}} \exp\Big\{-1/2(\boldsymbol{\beta}_j^{(c)}-\hat{\boldsymbol{\beta}_j^{(c)}})^T\Big[n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big](\boldsymbol{\beta}_j^{(c)}-\hat{\boldsymbol{\beta}_j^{(c)}})\Big\}\\
&\times\exp\Big\{-1/2\Big[\boldsymbol{\beta}_j^{(c)^T}[V_c\boldsymbol{I}]^{-1}\boldsymbol{\beta}_j^{(c)}\Big]\Big\}d\boldsymbol{\beta}_j^{(c)}\\
&= C_0\int_{\boldsymbol{\beta}_j^{(c)}} \exp\Big\{-\frac{1}{2}\Big[\boldsymbol{\beta}_j^{(c)^T}\Big(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})+(V_c\boldsymbol{I})^{-1}\Big)\boldsymbol{\beta}_j^{(c)}\\
&-2\boldsymbol{\beta}_j^{(c)^T}\Big(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big)\hat{\boldsymbol{\beta}_j^{(c)}}\\
&\quad+\hat{\boldsymbol{\beta}_j^{(c)}}^T\Big(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big)\hat{\boldsymbol{\beta}_j^{(c)}}\Big]\Big\}d\boldsymbol{\beta}_j^{(c)}\\
&= C_1\int_{\boldsymbol{\beta}_j^{(c)}} \exp\Big\{-\frac{1}{2}\Big[\Big(\boldsymbol{\beta}_j^{(c)}-\Sigma^*(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}))\hat{\boldsymbol{\beta}_j^{(c)}}\Big)^T\Sigma^{*-1}\\
&\quad\times\Big(\boldsymbol{\beta}_j^{(c)}-\Sigma^*(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}))\hat{\boldsymbol{\beta}_j^{(c)}}\Big)\Big]\Big\}d\boldsymbol{\beta}_i^{(c)},\\
&= C_1\Big[(2\pi)^{|E_j|/2}\Big]|\Sigma^*|^{1/2},
\end{aligned}
$$

where C_0 is a constant representing the normalizing constant for the prior of $\boldsymbol{\beta}_j^{(c)}$, C_1 is a constant combining C_0 and terms not involving $\boldsymbol{\beta}_j^{(c)}$, $\Sigma^* = \Big(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})+\frac{1}{V_c}\boldsymbol{I}\Big)^{-1}$, and $|E_j|$ is the number of parents of node i.

Recall that V_c is the variance in the prior distribution of $\boldsymbol{\beta}_j^{(c)}$ and chosen to be large to construct a non-informative but proper prior for $\boldsymbol{\beta}_j^{(c)}$. When the sample size n is large, information in the data dominates the priors,

$$
\begin{aligned}
|\Sigma^*|^{1/2} &= \Big|\Big(n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})+\frac{1}{V_c}\boldsymbol{I}\Big)\Big|^{-1/2}\\
&\asymp \Big|n\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big|^{-1/2}\\
&= n^{-|E_j|/2}\Big|\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big|^{-1/2}.
\end{aligned}
$$

We thus have

$$
\begin{aligned}
p(\boldsymbol{X},\boldsymbol{Y}|\eta=1,\boldsymbol{r}_j^{(c)}) &= p(\boldsymbol{Z}|\eta=1,\boldsymbol{r}^{(c)}) \\
&\approx \Big[(\frac{2\pi}{n})^{\sum_{i=1}^p |E_i|/2}\Big] \\
&\quad \prod_{j=1}^p \Big[L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big|\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})\Big|^{-1/2}\Big] \\
&\asymp C_z n^{-|E|/2}\prod_{j=1}^p L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)}),
\end{aligned}
$$

where $\sum_{j=1}^p |E_i| = |E|$. C_z is a constant, since $\Big|\bar{I}_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta = 1)\Big|^{-1/2}$ converges as $n \to \infty$ and based on assumption 3), $\Big[(2\pi/n)^{\sum_{j=1}^p |E_j|/2}\Big] \asymp \Big[n^{-\sum_{j=1}^p |E_j|/2}\Big]$. Note that $\prod_{j=1}^p L_j(\hat{\boldsymbol{\beta}_j^{(c)}}|\boldsymbol{Z_j},\eta=1,\boldsymbol{r}_j^{(c)})$ is a_η defined in equation (5) in the main text under the Bayesian context. In a Gibbs sampler, $\hat{\boldsymbol{\beta}_i^{(c)}}$ is represented by posterior samples of $\boldsymbol{\beta}_i^{(c)}$.

The same derivation applies to the calculation of $p(\boldsymbol{X},\boldsymbol{Y}|\eta=0,\boldsymbol{r}_j^{(x)},\boldsymbol{r}_j^{(y)})$, which gives

$$
\begin{aligned}
p(\boldsymbol{X},\boldsymbol{Y}|\eta=0,\boldsymbol{r}_j^{(x)},\boldsymbol{r}_j^{(y)}) &\approx C_{xy} n_x^{-|E_x|/2} n_y^{-|E_y|/2} \\
&\quad \times \prod_{j=1}^p \Big[L_j(\hat{\boldsymbol{\beta}_j^{(x)}}|\boldsymbol{X_j},\eta=0,\boldsymbol{r}_j^{(x)}) L_j(\hat{\boldsymbol{\beta}_j^{(y)}}|\boldsymbol{Y_j},\eta=0,\boldsymbol{r}_j^{(y)})\Big],
\end{aligned}
$$

where C_{xy} is constant, and, as above, $\prod_{i=j}^p \Big[L_j(\hat{\boldsymbol{\beta}_j^{(x)}}|\boldsymbol{X_j},\eta = 0,\boldsymbol{r}_j^{(x)}) L_j(\hat{\boldsymbol{\beta}_j^{(y)}}|\boldsymbol{Y_j},\eta = 0,\boldsymbol{r}_j^{(y)})\Big]$ is equivalent to b_η defined in equation (6) in the main text.

Now we have,

$$
\begin{aligned}
p(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) &= \frac{p(\boldsymbol{X}, \boldsymbol{Y}|(\cdot), \eta = 1)}{p(\boldsymbol{X}, \boldsymbol{Y}|(\cdot), \eta = 1) + p(\boldsymbol{X}, \boldsymbol{Y}|(\cdot), \eta = 0)} \\
&\approx \frac{C_z a_\eta(n^{-|E|/2})}{C_z a_\eta(n^{-|E|/2}) + C_{xy} b_\eta(n_x^{-|E_x|/2})(n_y^{-|E_y|/2})} \\
&= \Big[1 + \exp\{\log(b_\eta) - \log(a_\eta) + 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y) \\
&\quad + \log(C_{xy}/C_z)\}\Big]^{-1} \\
&= \Big[1 + \exp\{\log(b_\eta) - \log(a_\eta) + \lambda(n) + \log(C_{xy}/C_z)\}\Big]^{-1} \\
&\asymp \Big[1 + \exp\{\log(b_\eta) - \log(a_\eta) + \lambda(n)\}\Big]^{-1} \\
&= p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}),
\end{aligned}
$$

where $\lambda(n) = 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y)$. The last approximation is due to C_{xy}/C_z being bounded as $n \to \infty$, conditional on the following assumptions, 1) $|E|, |E_x|$, and $|E_y|$ are in the order of $O(p)$ and $|E| < |E_x| + |E_y|$, 2) n_x and n_y approaches to infinity in the same speed, and 3) $\log n_x/p \to \infty$ and $\log n_y/p \to \infty$ as $n_x, n_y, p \to \infty$. We denote the approximated conditional posterior of η as $p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$ with $\lambda(n)$ acting like a penalty determined by sample size and conditional on edges of inferred graphs.

Appendix II. Proof of the Proposition in Section 2.3.1

For any given ordering $\mathcal{O}$, let p denote the number of nodes, $|E_x|$ the number of edges in the network constructed based on data of sample size n_x from population X, $|E_y|$ the number of edges in the network based on data with size n_y from population Y, and $|E|$ the number of edges of the identical network constructed combining the two populations with sample size $n = n_x + n_y$.

$$
\begin{aligned}
p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) &\approx \Big[1 + \exp\{\log(b_\eta) - \log(a_\eta) + \lambda(n)\}\Big]^{-1} \\
\lambda(n) &= 1/2(|E| \log n - |E_x| \log n_x - |E_y| \log n_y),
\end{aligned}
$$

is the approximated conditional posterior probability for η.

Proposition: Assume 1) sparse networks with $|E|, |E_x|$, and $|E_y|$ in the order of $O(p)$, 2) $n_x \to \infty$ and $n_y \to \infty$ in the same speed, and 3) $\log n_x/p \to \infty$ as $n_x, p \to \infty$, and similar assumptions applied to n_y. Then $\lim_{n_x,n_y\to\infty} p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 1$ if the underlying $\eta = 1$, and $\lim_{n_x,n_y\to\infty} p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 0$ if the underlying $\eta = 0$.

Proof. We examine the property of $p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$ at the underlying values of η.

1. Underlying $\eta = 1$, i.e., the two populations share the same network. Set $n_x = c_1 n$ and $n_y = c_2 n$ with $0 < c_1, c_2 < 1$, we have

$$\begin{aligned} 1/2(|E_x|\log n_x + |E_y|\log n_y) &= 1/2\{|E_x|\log(c_1 n) + |E_y|\log(c_2 n)\} \\ &= 1/2(|E_x|\log n + |E_y|\log n + |E_x|\log c_1 \\ &\quad +|E_y|\log c_2). \end{aligned}$$

   For any given ordering, we assume $|E| < |E_x| + |E_y|$. That is, the two graphs have at least one edge in common and if an edge does not exist in each individual network, then it is not in the combined network either. We then have $1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y) = 1/2 \log n(|E| - |E_x| - |E_y|) - \log c_1|E_x|/2 - \log c_2|E_y|/2 \to -\infty$, as $n_x, n_y \to \infty$ (so does n). Furthermore, as $n_x, n_y \to \infty$, from the definitions of $\log a_\eta$ and $\log b_\eta$, $\log a_\eta - \log b_\eta \to 0$ as $n_x, n_y \to \infty$ when the underlying $\eta = 1$. Combining all these leads to

$$\log b_\eta - \log a_\eta + 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y) \to -\infty,$$

   which gives $p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) \to 1$ as $n_x, n_y \to \infty$ if the underlying $\eta = 1$.

2. Underlying $\eta = 0$, i.e., each of the two populations has its unique network

under a given ordering. Following the definition of a_η, we have

$$
\begin{aligned}
\log a_\eta &= \sum_{j=1}^{p}\Big\{ -\frac{n_x}{2}\log\sigma^2_{x(j)} - \frac{n_y}{2}\log\sigma^2_{y(j)} \\
&\quad -\frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(c)}\boldsymbol{X}_i)^T(\boldsymbol{X}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(c)}\boldsymbol{X}_i)}{2\sigma^2_{x(j)}} \\
&\quad -\frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(c)}\boldsymbol{Y}_i)^T(\boldsymbol{Y}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(c)}\boldsymbol{Y}_i)}{2\sigma^2_{y(j)}}\Big\} \\
&= \sum_{j=1}^{p}\Big\{ -\frac{n_x}{2}\log\sigma^2_{x(j)} - \frac{n_y}{2}\log\sigma^2_{y(j)} - \sum_{l_1=1}^{n_x}\frac{{\epsilon^c_{l_1}}^2}{2\sigma^2_{x(j)}} - \sum_{l_2=n_x+1}^{n}\frac{{\epsilon^c_{l_2}}^2}{2\sigma^2_{y(i)}}\Big\}, \\
\log b_\eta &= \sum_{j=1}^{p}\Big\{ -\frac{n_x}{2}\log\sigma^2_{x(j)} - \frac{n_y}{2}\log\sigma^2_{y(j)} \\
&\quad -\frac{(\boldsymbol{X}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(x)}\boldsymbol{X}_i)^T(\boldsymbol{X}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(s)}\boldsymbol{X}_i)}{2\sigma^2_{x(j)}} \\
&\quad -\frac{(\boldsymbol{Y}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(y)}\boldsymbol{Y}_i)^T(\boldsymbol{Y}_j - \sum_{i=1}^{j-1}\beta_{ij}^{(y)}\boldsymbol{Y}_i)}{2\sigma^2_{y(j)}}\Big\} \\
&= \sum_{j=1}^{p}\Big\{ -\frac{n_x}{2}\log\sigma^2_{x(j)} - \frac{n_y}{2}\log\sigma^2_{y(j)} - \frac{\sum_{l_1=1}^{n_x}{\epsilon^x_{l_1}}^2}{2\sigma^2_{x(j)}} - \frac{\sum_{l_2=n_x+1}^{n}{\epsilon^y_{l_2}}^2}{2\sigma^2_{y(j)}}\Big\}, \\
\log b_\eta - \log a_\eta &= \sum_{j=1}^{p}\Big(\frac{\sum_{l_1=1}^{n_x}({\epsilon^c_{l_1}}^2 - {\epsilon^x_{l_1}}^2)}{2\sigma^2_{x(j)}} + \frac{\sum_{l_1=n_x+1}^{n}({\epsilon^c_{l_2}}^2 - {\epsilon^y_{l_2}}^2)}{2\sigma^2_{y(j)}}\Big).
\end{aligned}
$$

In the following, property on one node is assessed and the results can be directly applied to the sum of all p nodes. If the underlying $\eta = 0$, that is, the relations among the nodes in the two populations are differential at least at one node, then, regardless of the ordering, forcing two differential networks to unify will result in larger random errors, i.e.,

$$
\begin{aligned}
({\epsilon^c_{l_1}}^2 - {\epsilon^x_{l_1}}^2)/n_x &\rightarrow {\sigma'}^2_{x(j)} - \sigma^2_{x(j)} > 0 \\
({\epsilon^c_{l_2}}^2 - {\epsilon^y_{l_2}}^2)/n_y &\rightarrow {\sigma'}^2_{y(j)} - \sigma^2_{y(j)} > 0,
\end{aligned}
$$

which lead to $\sum_{l_1=1}^{n_x}({\epsilon^c_{l_1}}^2 - {\epsilon^x_{l_1}}^2) \rightarrow \infty$ and $\sum_{l_1=n_x+1}^{n}({\epsilon^c_{l_2}}^2 - {\epsilon^y_{l_2}}^2) \rightarrow \infty$ in an ordering of $O(n)$ as $n_x, n_y \rightarrow \infty$, that is, $\log b_\eta - \log a_\eta \rightarrow \infty$ in an ordering of $O(n)$.

For $\lambda(n) = 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y)$ in the definition of $p^\lambda(\eta =$

$1|(\cdot), \boldsymbol{X}, \boldsymbol{Y})$,

$$\begin{aligned} & 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y) \\ = \quad & 1/2(|E| - |E_x| - |E_y|)\log n \\ & -1/2(|E_x|\log c_1 + |E_y|\log c_2) \\ = \quad & A\log n - 1/2(|E_x|\log c_1 + |E_y|\log c_2), \end{aligned}$$

where $A = 1/2(|E| - |E_x| - |E_y|)$. Since $|E| < |E_x| + |E_y|$, as $n_x, n_y \to \infty$, $A\log n \to -\infty$ in $O(\log n)$. Based on the sparsity assumption 1), $1/2(|E_x|\log c_1 + |E_y|\log c_2) \to \infty$ in the order p. Following assumption 3), we have $A\log n - 1/2(|E_x|\log c_1 + |E_y|\log c_2) \to \infty$ in $O(\log n)$, which is slower than $\log b_\eta - \log a_\eta \to \infty$ in an ordering of $O(n)$. Thus $\log b_\eta - \log a_\eta + 1/2(|E|\log n - |E_x|\log n_x - |E_y|\log n_y) \to \infty$, i.e., $p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) \to 0$ as $n_x, n_y \to \infty$.

In summary, $p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) \to 0$ as $n_x, n_y \to \infty$ when underlying $\eta = 0$ for any given ordering $\mathcal{O}$.

Combining results in 1. and 2. above, we have, for any given ordering $\mathcal{O}$, $\lim_{n_x, n_y \to \infty} p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 1$ if underlying $\eta = 1$, and $\lim_{n_x, n_y \to \infty} p^\lambda(\eta = 1|(\cdot), \boldsymbol{X}, \boldsymbol{Y}) = 0$ if underlying $\eta = 0$.

□

References

Almudevar A. A hypothesis test for equality of Bayesian network models. EURASIP Journal on Bioinformatics and Systems Biology 2010;2010:1–11.

Altomare D, Consonni G, La Rocca L. Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. Biometrics 2013;69(2):478–87.

Andersson SA, Madigan D, Perlman MD, et al. A characterization of markov equivalence classes for acyclic digraphs. The Annals of Statistics 1997;25(2):505–41.

Arshad SH, Holloway JW, Karmaus W, Zhang H, Ewart S, Mansfield L, Matthews S, Hodgekiss C, Roberts G, Kurukulaaratchy R. Cohort profile: The isle of wight whole population birth cohort (iowbc). International journal of epidemiology 2018;47:1043–4.

Ben-David E, Li T, Massam H, Rajaratnam B. High dimensional Bayesian inference for Gaussian directed acyclic graph models. arXiv preprint arXiv:11094371 2011;.

Van den Bergh M, Boix X, Roig G, de Capitani B, Van Gool L. Seeds: Superpixels extracted via energy-driven sampling. In: European conference on computer vision. Springer; 2012. p. 13–26.

Campos CPd, Ji Q. Efficient structure learning of Bayesian networks using constraints. Journal of Machine Learning Research 2011;12(Mar):663–89.

Canonne CL, Diakonikolas I, Kane DM, Stewart A. Testing bayesian networks. In: Conference on Learning Theory. PMLR; 2017. p. 370–448.

Cao X, Khare K, Ghosh M. Consistent bayesian sparsity selection for high-dimensional gaussian dag models with multiplicative and beta-mixture priors. Journal of Multivariate Analysis 2020;:104628.

Cao X, Khare K, Ghosh M, et al. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. The Annals of Statistics 2019;47(1):319–48.

Chickering DM. Learning equivalence classes of Bayesian-network structures. Journal of machine learning research 2002;2(Feb):445–98.

Consonni G, La Rocca L, Peluso S. Objective bayes covariate-adjusted sparse graphical model selection. Scandinavian Journal of Statistics 2017;44(3):741–64.

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

Eaton D, Murphy K. Bayesian structure learning using dynamic programming and mcmc. arXiv preprint arXiv:12065247 2012;.

Ellis B, Wong WH. Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association 2008;103(482):778–89.

Felix JF, Joubert BR, Baccarelli AA, Sharp GC, Almqvist C, Annesi-Maesano I, Arshad H, Baïz N, Bakermans-Kranenburg MJ, Bakulski KM, et al. Cohort profile: pregnancy and childhood epigenetics (pace) consortium. International journal of epidemiology 2017;47(1):22–23u.

Fernández C, Ley E, Steel M. Benchmark priors for bayesian model averaging. Journal of Econometrics 2001;100(2):381 – 427.

Friedman N, Koller D. Being Bayesian about network structure. a Bayesian approach to structure discovery in Bayesian networks. Machine learning 2003;50(1-2):95–125.

Fu F, Zhou Q. Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. Journal of the American Statistical Association 2013;108(501):288–300.

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

Giudici P, Green , PJ . Decomposable graphical Gaussian model determination. Biometrika 1999;86(4):785–801.

Han S, Wong RK, Lee TC, Shen L, Li SYR, Fan X. A full Bayesian approach for boolean genetic network inference. PloS one 2014;9(12):e115806.

Han S, Zhang H, Homayouni R, Karmaus W. An efficient Bayesian approach for Gaussian bayesian network structure learning. Communications in Statistics-Simulation and Computation 2016;doi:10.1080/03610918.2016.1143103.

Hastings WK. Monte carlo sampling methods using markov chains and their applications. Biometrika 1970;57:97–109.

Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning 1995;20(3):197–243.

Ishwaran H, Rao JS. Spike and slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics 2005;33:730–73.

Jacob L, Neuvial P, Dudoit S. More power via graph-structured tests for differential expression of gene networks. The Annals of Applied Statistics 2012;:561–600.

Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu CJ, et al. Dna methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. The American Journal of Human Genetics 2016;98(4):680–96.

Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun Ø, Cupul-Uicab LA, et al. 450k epigenome-wide scan identifies differential dna methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives 2012;120(10):1425.

Kuipers J, Moffa G. Partition mcmc for inference on acyclic digraphs. Journal of the American Statistical Association 2017;112:282–99.

Ladd-Acosta C, Shu C, Lee BK, Gidaya N, Singer A, Schieve LA, Schendel DE, Jones N, Daniels JL, Windham GC, et al. Presence of an epigenetic signature of prenatal cigarette smoke exposure in childhood. Environmental research 2016;144:139–48.

Larrañaga P, Poza M, Yurramendi Y, Murga RH, Kuijpers CMH. Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. IEEE transactions on pattern analysis and machine intelligence 1996;18(9):912–26.

Lee J, Chung W, Kim E, Kim S. A new genetic approach for structure learning of Bayesian networks: Matrix genetic algorithm. International Journal of Control, Automation and Systems 2010;8(2):398–407.

Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003;19:90–7.

Li H, Zhou Q. Gaussian DAGs on network data. 2019. `arXiv:1905.10848`.

Madigan D, Andersson SA, Perlman MD, Volinsky CT. Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. Communications in Statistics–Theory and Methods 1996;25(11):2493–519.

Madigan D, York J, Allard D. Bayesian graphical models for discrete data. International Statistical Review/Revue Internationale de Statistique 1995;:215–32.

Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. Journal of the American Statistical Association 1988;83:1023–32.

Moore A, Wong WK. Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In: ICML. volume 3; 2003. p. 552–9.

Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics 2012;4(2):199–203.

Ni Y, Müller P, Wei L, Ji Y. Bayesian graphical models for computational network biology. BMC bioinformatics 2018;19(3):63.

Oates CJ, Smith JQ, Mukherjee S, Cussens J. Exact estimation of multiple directed acyclic graphs. Statistics and Computing 2016;26(4):797–811.

Park YW, Klabjan D. Bayesian network learning via topological order. arXiv preprint arXiv:170105654 2017;.

Preiss B. Data structures and algorithms with object-oriented design patterns in C++. John Wiley & Sons, 2008.

Quraishi BM, Zhang H, Everson TM, Lockett GA, Ray M, Holloway JW, Arshad SH, Karmaus W. Identifying CpG sites associated with eczema via random forest screening of epigenome-wide DNA methylation. Journal of Allergy and Clinical Immunology 2015;135(2):AB158.

Rahman S, Khare K, Michailidis G, Martinez C, Carulla J. Estimation of Gaussian directed acyclic graphs using partial ordering information with an application to dairy cattle data. arXiv preprint arXiv:190205173 2019;.

Robert C. The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media, 2007.

Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika 2010;97(3):519–38.

Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. Journal of Econometrics 1996;75:317–43.

Squires C, Wang Y, Uhler C. Permutation-based causal structure learning with unknown intervention targets. Virtual: PMLR; volume 124 of *Proceedings of*

Machine Learning Research; 2020. p. 1039–48. URL: `http://proceedings.mlr.press/v124/squires20a.html`.

Städler N, Dondelinger F, Hill SM, Akbani R, Lu Y, Mills GB, Mukherjee S. Molecular heterogeneity at the network level: high-dimensional testing, clustering and a tcga case study. Bioinformatics 2017;33(18):2890–6.

Tsamardinos I, Brown LE, Aliferis CF. The Max-Min hill-climbing Bayesian network structure learning algorithm. Machine learning 2006;65(1):31–78.

Wang Y, Segarra S, Uhler C. High-dimensional joint estimation of multiple directed gaussian graphical models. arXiv preprint arXiv:180400778 2018;.

Wasserman L, Roeder K. High dimensional variable selection. Annals of statistics 2009;37(5A):2178.

Xia Y, Cai T, Cai TT. Testing differential networks with applications to the detection of gene-gene interactions. Biometrika 2015;:asu074.

Zellner A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel PK, Zellner A, editors. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam]; 1986. p. 233–43.

Zhang H, Huang X, Gan J, Karmaus W, Sabo-Attwood T, et al. A two-component g-prior for variable selection. Bayesian Analysis 2016;11(2):353–80.

Zhao SD, Cai TT, Li H. Direct estimation of differential networks. Biometrika 2014;101(2):253–68.

Zhou Q. Multi-domain sampling with applications to structural inference of Bayesian networks. Journal of the American Statistical Association 2011;106(496):1317–30.