Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

An efficient Bayesian approach for Gaussian Bayesian network structure learning

Shengtong Han, Hongmei Zhang, Ramin Homayouni & Wilfried Karmaus

To cite this article: Shengtong Han, Hongmei Zhang, Ramin Homayouni & Wilfried Karmaus (2017) An efficient Bayesian approach for Gaussian Bayesian network structure learning, Communications in Statistics - Simulation and Computation, 46:7, 5070-5084, DOI: 10.1080/03610918.2016.1143103

To link to this article: http://dx.doi.org/10.1080/03610918.2016.1143103

Accepted author version posted online: 18 Feb 2016.
Published online: 18 Feb 2016.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 7, 5070–5084
http://dx.doi.org/10.1080/03610918.2016.1143103

Taylor & Francis
Taylor & Francis Group

An efficient Bayesian approach for Gaussian Bayesian network structure learning

Shengtong Han, Hongmei Zhang, Ramin Homayouni, and Wilfried Karmaus

School of Public Health, Bioinformatics Program and Center for Translational Informatics, University of Memphis, Memphis, TN, United States

ABSTRACT

This article proposes a Bayesian computing algorithm to infer Gaussian directed acyclic graphs (DAGs). It has the ability of escaping local modes and maintaining adequate computing speed compared to existing methods. Simulations demonstrated that the proposed algorithm has low false positives and false negatives in comparison to an algorithm applied to DAGs. We applied the algorithm to an epigenetic dataset to infer DAG's for smokers and nonsmokers.

ARTICLE HISTORY
Received 19 May 2015
Accepted 11 January 2016

KEYWORDS
DNA methylation; Gaussian DAG; MCMC

MATHEMATICS SUBJECT CLASSIFICATION
62; 62H12

1. Introduction

There is an increasing attention on learning directed acyclic graphs (DAG's), or Bayesian networks, which have an appealing property of encoding conditional independence relations by graphs explicitly. By definition, a graph is a DAG if all the links (edges) are directed, but there are no directed loops (circles). Conditional independence means each variable in the graph is conditionally independent of its nondescendants given its parent set. Two DAGs are Markov equivalent if they represent the same set of conditional independence. To characterize the class of DAGs with the same set of conditional independence, a partially directed acyclic graph (CPDAG) is introduced, which may be with both directed links and undirected links. One can refer to Andersson et al. (1997), Chickering (2002) for details. It still remains a challenge for inferring Bayesian networks due to network complexity and the existence of multiple local maximums. A number of algorithms are devoted to estimate Bayesian networks, including greedy local search (Heckerman and Chickering, 1995), Optimal reinsertion search (Moore and keen Wong, 2003), Max–Min Hill-Climbing (Tsamardinos et al., 2006), genetic algorithm (Larranaga et al., 1996; Lee et al., 2010), dynamic programming (Eaton, 2007), branch-and-bound algorithm (de Campos et al., 2011), Markov Chain Monte Carlo (MCMC) approaches (Ellis and Wong, 2008; Friedman and Koller, 2003; Giudici and Green, 1999; Han et al., 2014; Madigan et al., 1995, 1996; Zhou, 2011). The PC algorithm, proposed by Spirtes et al. (2000), becomes very popular to infer DAGs by detecting the conditional independence, then orienting the links as long as there are no directed circles in the resulting network. It is worthy noting that the PC algorithm finds completed partially directed acyclic graphs, not single DAGs. Subsequently, related work include Harris and Drton (2013), Kalisch et al. (2007), and Kalisch and Bhlmann (2008).

CONTACT Hongmei Zhang hzhang6@memphis.edu School of Public Health, Bioinformatics Program and Center for Translational Informatics, University of Memphis, Memphis, TN, United States.

Among these, Friedman and Koller (2003) made an important improvement in inferring Bayesian networks by introducing graph order MCMC and previous work include Bouckaert (1994). Assuming the order is known, Shojaie and Michailidis (2010) proposed an efficient penalized likelihood method for the estimation of the adjacent matrix of directed graphs; Altomare et al. (2013) proposed an objective method for DAG inference. Even with order MCMC, it is still difficult to capture the underlying distributions because of the common problem of multiple local maximums, as experimentally demonstrated by Ellis and Wong (2008). To improve sampling efficiency, Ellis and Wong (2008) suggested the use of advanced sampling technique-single queue Equi-Energy sampler, a variant of the Equi-Energy sampler proposed by Kou et al. (2006) to obtain more reliable posterior samples.

However, most efforts, although stated applicable to continuous variables, focus on discrete networks in which each node is a categorical variable such as in the work by Ellis and Wong (2008), Liang and Zhang (2009), Zhou (2011), and Balov (2013), and among others. In biomedical studies, to utilize these methods, continuous variables have to be transformed into discrete variables. This discretization manipulation may induce the risk of losing some meaningful information, for instance, the dependence relationships among variables. Furthermore, some datasets are continuous in nature and discretizing them may practically be meaningless. In this work, these limitations motivate us to investigate the structure of Bayesian networks from continuous data.

Section 2 presents the model, including an introduction to Bayesian network, priors, and posteriors for parameters. The posterior computing, e.g., the proposed algorithm, is given in Section 3. Simulation studies and a real-data application are included in Section 4. It ends with summary and discussion in Section 5.

2. The model

Bayesian network is a directed acyclic graph which can be characterized by two components, a set of nodes $\boldsymbol{X}$ (random variables), and a collection of directed links, $\boldsymbol{E}$. Throughout we assume there are n nodes, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, in the network, and the data are fully observed without any missing values. The joint distribution of $\boldsymbol{X}$ satisfies, by Markov property

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i|Pa(X_i)),$$

where $Pa(X_i) \subset \boldsymbol{X} \setminus \{X_i\}$ is the parent set of X_i. We assume each X_i is a Gaussian random variable with its mean as a function of its parents, i.e., $X_i = \beta_{i0} + \sum_{j:X_j \in Pa(X_i)} \beta_{ij} X_j + \epsilon_i$, $i = 1, 2, \ldots, n$. Parameters β_{ij} are coefficients and ϵ_i is assumed to be Gaussian random noise with mean 0 and variance σ_i^2. We have

$$X_i|Pa(X_i) \sim N\left(\mu_{X_i}, \sigma_i^2\right),$$

where $\mu_{X_i} = \beta_{i0} + \sum_{j:X_j \in Pa(X_i)} \beta_{ij} X_j$. Let $\boldsymbol{\theta_i}$ be the set of parameters associated with node X_i, i.e., $\boldsymbol{\theta_i} = \{\beta_{i0}, \beta_{i1}, \beta_{i2}, \ldots, \beta_{iT_i}, \sigma_i^2\}$ and all parameters for the network are denoted by $\boldsymbol{\theta} = \cup_{i=1}^{n} \boldsymbol{\theta_i}$, where $T_i = |Pa(X_i)|$, the number of parents for node X_i. In the following, we assume $T_i \leq 4$, i.e., the maximum number of parents is no more than 4. The assumption

of boundness on the number of parents is quite common and unavoidable in Bayesian network structure learning since graphs with overwhelmingly large number of links are usually preferred, resulting in overfitting problem (Ellis and Wong, 2008; Fu and Zhou, 2013; Zhou, 2011).

Suppose there are M independent observations for each variable in the network, $x_{ih}, i = 1, 2, \ldots, n; h = 1, 2, \ldots, M$. Let $\mathbb{X} = \{x_{ih}, i = 1, 2, \ldots, n; h = 1, 2, \ldots, M\}$. Given the graph $\mathcal{G}$ and the parameter associated with this graph, $\boldsymbol{\theta}^{\mathcal{G}}$ ($\boldsymbol{\theta}^{\mathcal{G}}$ has the same definition as $\boldsymbol{\theta}$ but is specific to graph $\mathcal{G}$), we have

$$P(\mathbb{X}|\mathcal{G}, \boldsymbol{\theta}^{\mathcal{G}}) = \prod_{h=1}^{M} \prod_{i=1}^{n} P(x_{ih}|Pa(x_{ih}), \boldsymbol{\theta_i}^{\mathcal{G}}).$$

Prior Distributions

To infer the parameter $\boldsymbol{\theta}^{\mathcal{G}}$, we use the fully Bayesian approach. We choose inverse gamma for the prior distribution of σ_i^2, i.e., $\sigma_i^2|\delta_i, \psi_i \sim \text{Inv-Gamma}(\delta_i, \psi_i)$, with δ_i, ψ_i known and set the conditional priors for $\boldsymbol{\beta_i} = (\beta_{i0}, \beta_{i1}, \beta_{i2}, \ldots, \beta_{iT_i})^T$ to be of the form $\boldsymbol{\beta_i}| \sim \text{MVN}(0, \sigma_i^2(\boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i^{Pa}})^{-1})$, where "MVN" denotes the multivariate normal distribution, $\boldsymbol{X_i^{Pa}}$ is an $M \times (T_i + 1)$ matrix and represents the observational data for the parents of node X_i. This prior allows us to obtain the posterior in the closed form in (1).

The prior distribution over structures is commonly chosen to be the uniform distribution (Friedman and Koller, 2003; Heckerman, 1999), and the sparse network is usually preferred. In our approach, similar to Ellis and Wong (2008), we consider structure priors as a function of the number of links as $P(\mathcal{G}) \propto \gamma^{\sum_{i=1}^{n} |Pa^{\mathcal{G}}(X_i)|}$, for some $0 < \gamma < 1$.

Posterior Distributions

The joint posterior distribution of $(\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G})$ is

$$P(\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G}|\mathbb{X}) \propto P(\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G}) \times P(\mathbb{X}|\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G})$$

By integrating out the parameters, the marginal posterior distribution of $\mathcal{G}$ is

$$P(\mathcal{G}|\mathbb{X}) \propto \prod_{i=1}^{n} \gamma^{T_i} \left[\pi^{\frac{T_i+1}{2}} \left| \boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i^{Pa}} \right|^{-\frac{1}{2}} \times \frac{\Gamma\left(\frac{M}{2} + \delta_i\right)}{\left\{ \psi_i - \frac{1}{4} \left(\boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i} \right)^T \boldsymbol{\eta_i} + \frac{1}{2} \boldsymbol{X_i}^T \boldsymbol{X_i} \right\}^{\frac{M}{2}+\delta_i}} \right], \quad (1)$$

where $\boldsymbol{X_i^{Pa}}$ is an $M \times (T_i + 1)$ matrix with the hth row $(1, x_{h1}^{(i)}, \ldots, x_{hT_i}^{(i)})$, $h = 1, 2, \ldots, M$, $\boldsymbol{X_i} = (x_{i1}, x_{i2}, \ldots, x_{iM})^T$, $i = 1, 2, \ldots, n$, an $M \times 1$ observation vector for node X_i, and $\boldsymbol{\eta_i} = (\boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i^{Pa}})^{-1} \boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i}$. The derivation is given in the Appendix. When X_i has no parent, i.e. $|T_i| = 0, \boldsymbol{X_i^{Pa}} = (1, 1, \ldots, 1)^T, \boldsymbol{X_i} = (x_{i1}, x_{i2}, \ldots, x_{iM})^T$. Thus

$$P(X_i; Pa(X_i) = \emptyset) = \sqrt{\frac{\pi}{M}} \frac{\Gamma\left(\frac{M}{2} + \delta_i\right)}{\left[\psi_i - \frac{\left(\sum_{h=1}^{M} x_{ih}\right)^2}{4M} + \frac{1}{2} \sum_{h=1}^{M} x_{ih}^2 \right]^{\frac{M}{2}+\delta_i}}.$$

3. Posterior computing

3.1. Adjusted single queue equi-energy sampler (ASQEE)

Many algorithms for network structure inference are available, as described in Section 1. Learning Bayesian networks is known to be NP hard (Chickering et al., 2004) and it poses great challenge on many traditional learning algorithms using MCMC on individual network structures (Giudici and Green, 1999; Grzegorczyk and Husmeier, 2008; Madigan et al., 1995, 1996). Another difficulty in network inference stems from the nonregular posterior distribution of the networks of interest. Friedman and Koller (2003) made a substantial improvement by focusing on graph orders (i.e., treating graph ordering as a random variable), rather than on individual network structures, to build a Markov chain since the space of orders is experimentally verified to be much smaller and more regular than the space of structures. Instead of inferring graphs using (1), we follow the idea of Friedman and Koller (2003) to estimate the orders and construct the graphs based on inferred linked edges.

Given a directed acyclic graph, there exists at least one total ordering, $\mathcal{O}$ (an ordering of n variables, $X_1, X_2, \ldots, X_n$), in which X_i proceeds X_j if $X_i \in Pa(X_j)$. On the other hand, if X_i proceeds X_j in order $\mathcal{O}$, directed links from X_j to X_i are prohibited in all of its consistent graphs, imposing an restriction on network structures. One advantage of graph orders is that it is not necessary to perform acyclicity checking, which is commonly required when updating individual network by edge addition, deletion, or reversal. In Friedman and Koller (2003), a Markov chain on graph orders $\mathcal{O}$ and its posterior probability is obtained using all consistent graphs,

$$P(\mathcal{O}|\mathbb{X}) \propto P(\mathcal{O}) \times \sum_{\mathcal{G}:\mathcal{G}^{\mathcal{O}}} P(\mathbb{X}|\mathcal{G}, \mathcal{O}) P(\mathcal{G}|\mathcal{O}),$$

where $\mathcal{G}^{\mathcal{O}}$ are all consistent graphs with order $\mathcal{O}$. However, the enumeration of all consistent graphs becomes practically intractable for large networks. Instead of sampling all consistent graphs, Ellis and Wong (2008) suggested sampling a number of distinct consistent graphs such that

$$\sum_{i=1}^{k} P(\mathcal{G}_i|\mathcal{O}) > 1 - \epsilon,$$

where ϵ is a prespecified small positive number, $\mathcal{G}_i$, $i = 1, 2, \ldots, k$ are distinct consistent graphs. However, $P(\mathcal{G}_i|\mathcal{O})$ is usually known up to a normalization constant, and calculation of the exact probability is practically impossible. To bypass this difficulty, given an order, we propose to sample a number of consistent graphs and only use the one with highest probability.

Even with smaller space of orders, the problem of local maximum is still severe. To avoid being trapped at local maximums when drawing posterior samples of $\mathcal{O}$, we propose to use the Single-Queue Equi-Energy sampling (SQEE) proposed by Ellis and Wong (2008), a variant of the Equi-Energy (EE) sampler (Kou et al., 2006). The EE sampling proceeds as follows. The first step is to define a sequence of temperatures $1 = T_1 < T_2 < \cdots < T_W$, where W is the number of chains. Then, a sequence of energy levels are introduced, i.e., $H_1 < H_2 < \cdots < H_W$ where $H_1 \leq \min_x H(x)$, $H_W \leq \infty$. Based on these energy levels, the tempered distributions are defined as $\pi_l(x) = \exp(\frac{-\max\{H(x), H_l\}}{T_l})$ $l = 1, 2, \ldots, W$, which is the target distribution of lth chain and the energy rings are constructed as $D_l = \{x|H(x) \in [H_l, H_{l+1}), l = 1, 2, \ldots, W\}$. By definition, the larger the value of l, the more flatten distribution the lth chain

gets, enhancing the ability of the chain jumping across different modes. Specifically, when $l = 1$, $\pi_1(x)$ is the target distribution and $\pi_1(x) = \exp(-H(x))$. This illustrates the relation between energy function and the distribution. That is $H(x) = -\log(\pi_1(x)) = -\log(P(\mathcal{O}))$. One can refer to Kou et al. (2006) for detailed discussions. The SQEE sampler is the same as the EE sampler, except for the sampling of orders at each chain. The construction of the lth chain in the EE sampler is only based on the $(l+1)$th chain. The SQEE sampler, on the other hand, uses information from any of previous higher order chains, $l+1, l+2, \ldots, W$, which makes a better communications between different chains.

Due to the use of a single graph in the SQEE sampler for a given order, we name the sampler as the adjusted single queue equi-energy sampler (ASQEE). Admittedly, the use of only one representative graph with highest posterior probability may result in bias in calculating the posterior probability of an order. However, it significantly reduces the computing difficulty in Ellis and Wong (2008). Our simulations discussed later indicate that this approximation is applicable and produces reasonably good result.

3.2. Ordering proposal

As discussed in earlier sections, graph ordering is a random variable. Order proposal plays a crucial role in building the Markov chain. We still follow the "cylindrical shift" operation (Ellis and Wong, 2008). To improve the convergence of Markov chain, the number of nodes to be flipped in proposing orders is adjusted in a dynamic way for each single chain. Specifically, the number of flipping nodes decreases as the number of iterations grows. That is, at the beginning of MCMC, we expect the chain to move fast from the current stage, while when the MCMC reaches a relatively stable state, then we expect it to move forward slowly. This will help the MCMC chain reach a more reliable stationary distribution quickly. 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. Thus, the graph order is random and to be updated in the MCMC. The detailed algorithm is presented in Algorithm 1.

3.3. Performance evaluation

Our interest is on directed links of the network, e.g., the link $X_i \to X_j$, and the associated posterior probabilities. Given a collection of a number of order samplers, $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_K$, the posterior probability of a directed link f can be calculated

$$\begin{aligned} P(f|\mathbb{X}) &= \sum_k P(f, \mathcal{O}_k|\mathbb{X}) \\ &= \sum_k P(\mathcal{O}_k|\mathbb{X}) \times P(f|\mathcal{O}_k, \mathbb{X}) \\ &= \sum_k \left[P(\mathcal{O}_k|\mathbb{X}) \times \sum_i P(f|\mathcal{G}_i^{\mathcal{O}_k}\mathcal{O}_k, \mathbb{X}) P(\mathcal{G}_i^{\mathcal{O}_k}|\mathcal{O}_k, \mathbb{X}) \right], \end{aligned} \tag{2}$$

where $\mathcal{G}_i^{\mathcal{O}_k}$ are consistent graphs sampled for order $\mathcal{O}_k$. By using (2), for a set of links of interest, the posterior probabilities can be calculated. Given a threshold, a positive number between 0 and 1, a set of directed edges could be collected with posterior probability greater than the

Algorithm 1 Adjusted Single Queue Equi-Energy algorithm (ASQEE)

Assign $X_1^{(W)}$ an initial ordering, set $\widehat{D}_l = \emptyset, l = 1, 2, \ldots, W$.
For $n = 1, 2, \ldots$
For $l = W, W-1, \ldots, 1$

if $n > (W-l)(B+N)$ (B is the burn in period) **then**
  do
  if $l = W$ or if $\widehat{D}_{I(X_{n-1}^{(i)})} = \emptyset$ **then**
    Perform local M-H move to update $X_{n-1}^{(l)}$ by $X_n^{(l)}$ with target distribution π_l
  else if $l < W$ AND if $\widehat{D}_{I(X_{n-1}^{(l)})} \neq \emptyset$ **then**
    Generate $\mu \sim U(0, 1)$
    if $\mu > p_{ee}$ **then**
      Perform local M-H move to update $X_{n-1}^{(l)}$ by $X_n^{(l)}$ with target distribution π_l
    else if $\mu \leqslant p_{ee}$ **then**
      Uniformly pick a state y from $\widehat{D}_{I(X_{n-1}^{(l)})}$ ($\widehat{D}_{I(X_{n-1}^{(l)})}$ is the union of all previous energy rings with similar energy level) and $X_n^{(l)} \longleftarrow y$ with probability $\min(1, \frac{\pi_1(y)Q(y;X_{n-1}^{(l)})}{\pi_1(X_{n-1}^{(l)})Q(X_{n-1}^{(l)};y)})$ ($Q(.,.)$ is the transition kernel function); $X_n^{(l)} \longleftarrow X_{n-1}^{(l)}$ with the remaining probability
    end if
  end if
  if $n > (W-l)(B+N) + B$ **then**
    $\widehat{D}_{I(X_n^{(l)})} \longleftarrow \widehat{D}_{I(X_n^{(l)})} + \{X_n^{(l)}\}$
  end if
end if

threshold. Furthermore, false positives and false negatives can be obtained as well for different threshold for the purpose of comparing between different methods in Bayesian network inferences.

In simulations, $P(\mathcal{O}_k|\mathbb{X})$ is estimated using its relative frequency over all sampled orders after burn in. Similarly, we estimated $P(f, \mathcal{G}_i^{\mathcal{O}_k}|\mathcal{O}_k, \mathbb{X})$ by using the relative frequency of related consistent graphs over all sampled consistent graphs of the order.

4. Numerical studies

4.1. Simulated experiments

Simulation scenarios. In this section, networks with 10 nodes are considered. An order of these 10 nodes is generated via random rearrangement of $r_1, r_2, \ldots, r_{10}$ with $r_i \in \{1, 2, \ldots, 10\}$, i.e., $\mathcal{O}^* : r_1, r_2, \ldots, r_{10}, r_i \in \{1, 2, \ldots, 10\}$. Generation of a consistent graph proceeds in the following way, similar to Ellis and Wong (2008). For each node r_i, select the number of its parents, n_i, between 0 and $\min\{(i-1), R\}$, R is the maximum parents allowed, then assign n_i parents to node r_i chosen from $i-1$ proceeding nodes. The nodes and the parents of each node consist of the whole network structure. The network structure used to simulated datasets is presented in Fig. 1. Given this order and network structure, multiple datasets are generated with the following parameters: $\epsilon_i \sim N(0, \sigma_i^2)$; $\boldsymbol{\beta_i} \sim MVN(\boldsymbol{\mu}_{\boldsymbol{\beta}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_i})$; $\boldsymbol{\mu}_{\boldsymbol{\beta}_i} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a $(T_i+1) \times 1$ vector with all entries being 2, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_i}$ are both diagonal matrix with

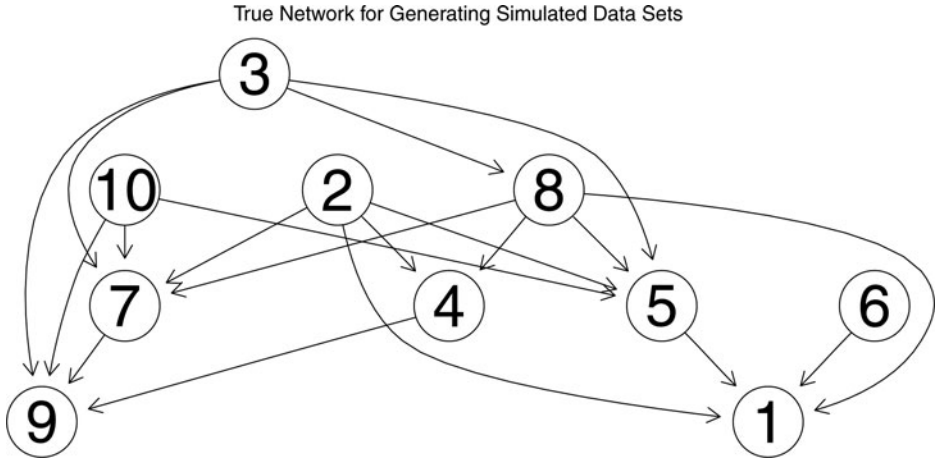

Figure 1. True network structure used for generating simulated datasets.

all elements being 0.1. To evaluate the effect of sample size, we set $M = 100, 500, 1000$. The impact of variance of the inference of network is also considered; σ_i is considered at two levels, $\sigma_i = 0.5, 1.0$.

To the best of our knowledge, FK algorithm (Friedman and Koller, 2003) is the only order based sampling approach which can make a fair comparison with the proposed ASQEE. Another Bayesian approach by Altomare et al. (2013) focuses on the situation with the order known, which is different from our case. Both FK and ASQEE are tested on generated datasets. To implement FK, an initial order is randomly generated and the total number of iterations is set at 20,000 with the first 10,000 as the burn-in period. To reduce the auto-correlations within the chain, order samples are collected every 200 iterations after burn-in. Thus, 50 order samples are collected in the final analysis. To make the result from ASQEE comparable to that from FK, the top 50 order samples from ASQEE in terms of posterior probability are used to make the final inference.

Result. We applied both the ASQEE and FK algorithm on 10 simulated datasets generated following each of the simulation scenarios. The averaged number of false positive and false negatives with different parameter settings is reported in Figs. 2 and 3, respectively. For the ASQEE approach, at a given level of noise, increasing sample size will result in better inference as indicated by smaller number of false positives and false negatives (Figs. 2(d) and 3(d)). Higher level in noise seemed to weaken the quality of the inferences, but they are improved

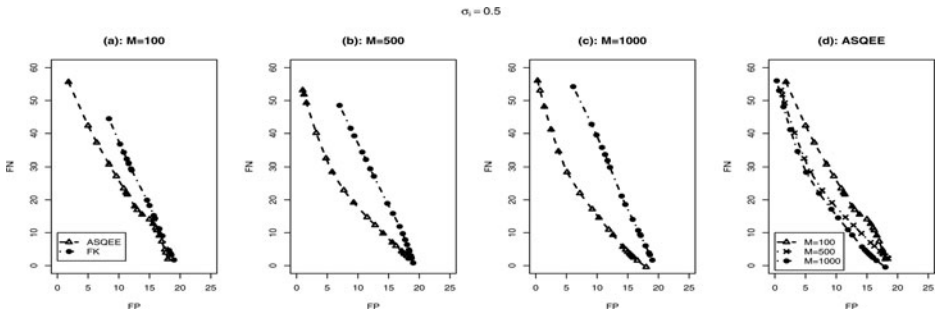

Figure 2. Comparison between ASQEE and FK ((a), (b), (c)) and sample size effect of ASQEE (d) in terms of average number of false positives and false negatives across 10 randomly generated datasets with $\sigma_i = 0.5$, $M = 100, 500, 1000$.

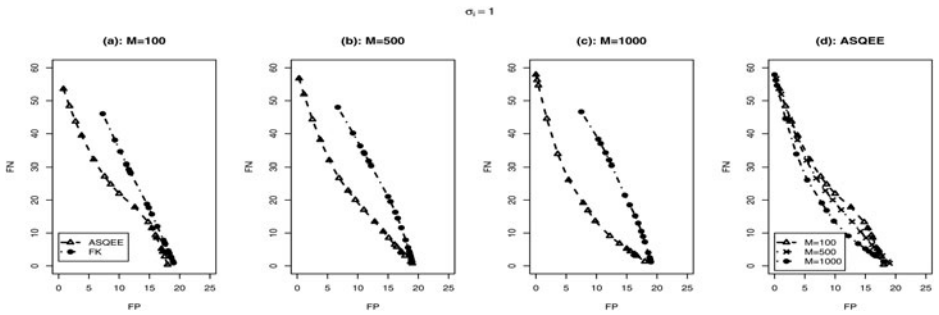

Figure 3. Comparison between ASQEE and FK ((a), (b), (c)) and sample size effect of ASQEE (d) in terms of average number of false positives and false negatives across 10 randomly generated data sets with $\sigma_i = 1.0$, $M = 100, 500, 1000$.

as the sample size gets larger. However, for the FK method, in general, it has higher false positives and false negatives. This may be due to its weak ability of escaping local modes. Overall, regardless of the sample size and the level of noise, ASQEE performs better than FK in terms of the averaged number of false positives and false negatives in both scenarios (Figs. 2(a)–(c) and 3(a)–(c)).

The findings are as expected. ASQEE has a better ability of escaping local maximum traps but the FK does not. Our simulations also indicate that using all consistent graphs result in a smaller bias compared to using one single consistent graph in calculating the posterior probability of an order. For small networks ($n < 10$) where enumerating all consistent graphs is possible, it is still recommended to use as many consistent graphs as possible. However, for large networks ($n \geq 10$), selecting one consistent graph with highest posterior probability using M-H is likely to be a good choice.

To further test the better ability of the proposed method in identifying true links, we simulated datasets based the graph in Fig. 1, but added two additional noisy nodes generated from normal distribution with mean 0 and variance 0.25. These two nodes are isolated, i.e., without any links to the true graph or between each other. For the purpose of demonstration, we focus on the case where $M = 100$, $\sigma_i = 0.5$. Small-sample size ($M = 100$) is considered since when sample size is large, the impact of noisy nodes is expected to be reduced. Given a threshold within interval (0, 1), a set of directed links are collected with the posterior probability higher than the cutoff. Based on the collection of directed links, a node is said to be present if it is in the collection. Then, we calculated the relative frequency of each node in the collection across 10 randomly generated datasets. The averages of relative frequency are displayed in Fig. 4 for the true nodes and noisy nodes, respectively. Consistent low frequencies indicate the method is robust with respect to noisy nodes.

Large network inference. The above simulations are based on networks with 10 nodes. We further applied the ASQEE to networks with 50 nodes to assess its ability in dealing with larger networks. The simulation procedure follows the same way as previously stated except that we included four times more nodes in the networks. The averaged numbers of false positives and false negatives of ASQEE and FK are graphically presented in Figs. 5 and 6. The FK approach never reached high false positives and never reached low false positives either. On the other hand, the proposed approach (ASQEE) gives a much longer span on possible false positives and false negatives compared to the FK approach, emphasizing the weak ability of FK escaping local modes as noted earlier. Among comparable false positives and false negatives (the lower right corner of each plot in Figures 5 and 6), apparently, our approach overall outperforms FK by observing lower false positives and false negatives.

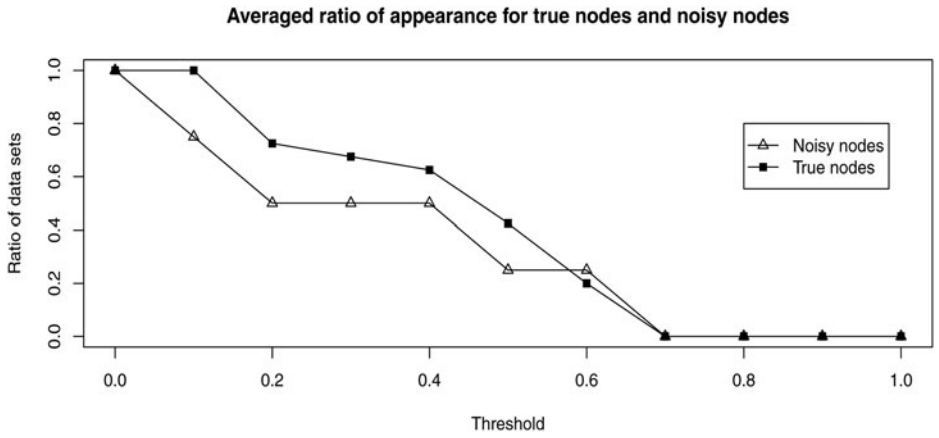

Figure 4. Averaged ratio of appearance for both true nodes and noisy nodes across 10 randomly generated datasets with $M = 100$, $\sigma_i = 0.5$.

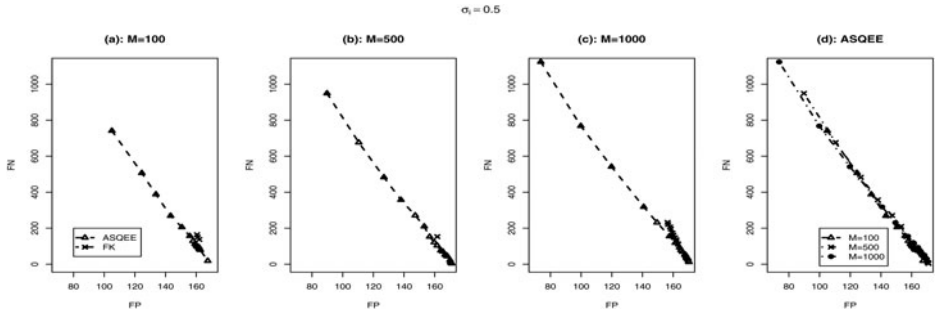

Figure 5. Comparison between ASQEE and FK ((a)–(c)) and sample size effect of ASQEE (d) in terms of average number of false positives and false negatives across 10 randomly generated datasets with $\sigma_i = 0.5$, $M = 100, 500, 1000$ for larger network with 50 nodes.

Moreover, the computing time of FK is much longer than that of ASQEE. This is as expected, since the number of networks is growing exponentially as the number of nodes becomes larger and the FK approach estimates exact posterior probabilities instead of highest posterior probability as in the proposed method. Compared to the performance when dealing with smaller networks with 10 nodes, both ASQEE and FK do not perform equally well when

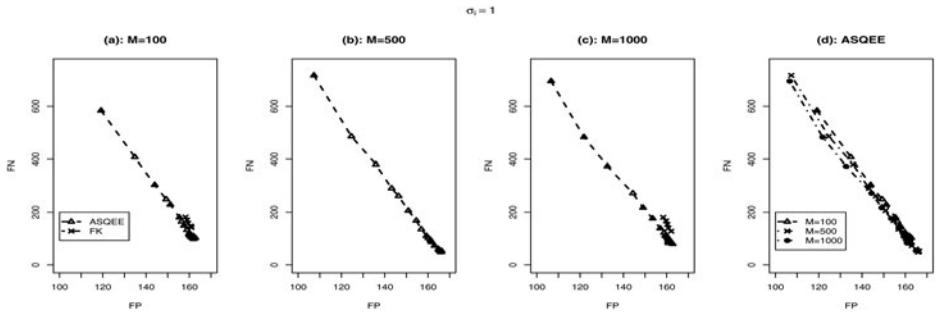

Figure 6. Comparison between ASQEE and FK ((a)–(c)) and sample size effect of ASQEE (d) in terms of average number of false positives and false negatives across 10 randomly generated datasets with $\sigma_i = 1.0$, $M = 100, 500, 1000$ for larger network with 50 nodes.

working with larger networks, indicating a great need of developing inference algorithms for large networks.

4.2. An application to A DNA methylation data set

To illustrate the proposed method, we apply it to a set of 26 CpG sites in 10 genes which are related to maternal smoking (Joubert et al., 2012) and aim to explore the connections among them using networks. The relevant information for all 26 CpG sites are given in Table 1. DNA methylation data of 245 girls measured at age 18 is used in the analysis. These 245 subjects are a random sample from the Isle of Wight birth cohort (Arshad and Hide, 1992). Among these 245 girls, 48 were exposed to maternal smoking during pregnancy.

To apply the proposed algorithm to the methylation dataset, logit transformation is applied to transform the methylation data in the interval $(0, 1)$ to the whole real line domain. The multivariate assumption of the logit-transformed DNA methylation data was satisfied based on Shapiro–Wilk normality tests. We applied the ASQEE algorithm to the data of 197 subjects not exposed and to the 48 exposed subjects. Five chains are run sequentially from high order to low, each with the total number of 3000 iterations and the first 2700 iterations are treated as burn in. Order samples are collected for final analysis. Given an order, Metropolis-Hastings algorithm is applied to sample a number of consistent graphs and the one with highest posterior probability is used to construct the pool of directed links, each with its posterior probability estimated by the frequency of the graph sampled in MCMC after convergence.

With different cutoffs on posterior probability, a varying number of directed links are obtained. So, the choice of cutoffs becomes crucial in building the finally inferred network. We follow the rule that a cutoff should be chosen such that the resulting network is of moderate

Table 1. Relevant information on 26 CpG sites. Chr denotes the chromosome location.

Chr	Gene	CpG	CpG Index in the inferred network
1	GFI1	cg10399789	10
1	GFI1	cg09662411	8
1	GFI1	cg06338710	7
1	GFI1	cg18146737	18
1	GFI1	cg12876356	15
1	GFI1	cg18316974	19
1	GFI1	cg09935388	9
1	GFI1	cg14179389	16
5	AHRR	cg23067299	25
5	AHRR	cg03991871	2
5	AHRR	cg05575921	6
5	AHRR	cg21161138	22
6	HLA-DPB2	cg11715943	11
7	MYO1G	cg19089201	21
7	MYO1G	cg22132788	23
7	MYO1G	cg04180046	3
7	MYO1G	cg12803068	14
7	ENSG00000225718	cg04598670	4
7	CNTNAP2	cg25949550	26
8	EXT1	cg03346806	1
14	TTC7B	cg18655025	20
15	CYP1A1	cg05549655	5
15	CYP1A1	cg22549041	24
15	CYP1A1	cg11924019	12
15	CYP1A1	cg18092474	17
21	RUNX1	cg12477880	13

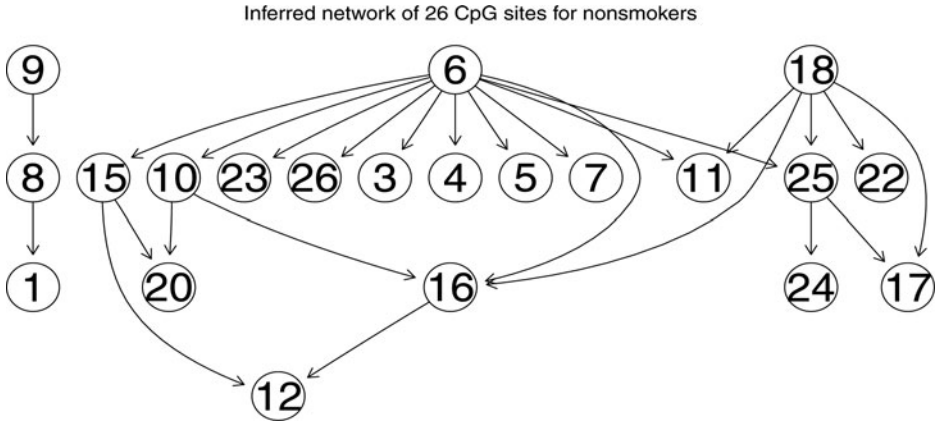

Figure 7. This inferred network is for 197 girls exposed to mother smoking during pregnancy with cutoff 0.25. CpG sites indices are given in Table 1.

size ("moderate size" means most of the nodes appearing and the number of links is approximately two or three times of the number of nodes) and captures the important information on the connections between genes. The finally inferred graph is then built upon the selected directed links. The result networks for nonsmokers and smokers are presented in Figs. 7 and 8, respectively, inferred based on similar thresholds.

The network inferred for the nonsmokers (Fig. 7) indicates a strong regulatory path from cg 05575921 (CpG index number 6; Table 1) in AHRR gene to CpG sites in genes GFI1, MYO1G, CNTNAP2, ENSG00000225718, CYP1A1, and HLA-DPB2. This is consistent with a recent finding on the AHRR gene related to its potential as a biomarker for smoking (Philibert et al., 2013). The dominance of cg 05575921 does not change in the network for smokers for cg 06338710 (index number 7) in gene GFI1 and cg 04180046 (index number 3) in gene MYO1G. Although the direct control of cg05575921 to some CpG sites in the network for nonsmokers disappeared in the network for smokers (Fig. 8), indirect connection is still observed for most CpG sites shown in Fig. 7. For instance, the regulatory function of cg05575921 to CpG sites cg10399789 (index number 10) and cg 12876356 (index number 15)

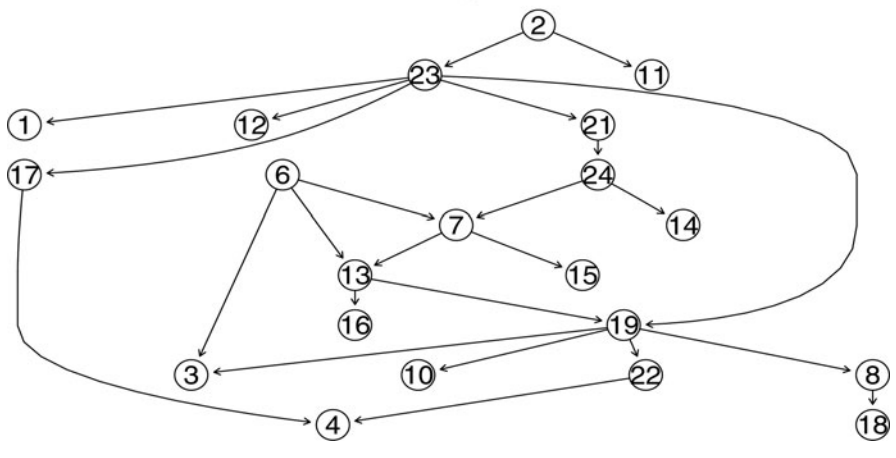

Figure 8. This inferred network is for girls of 48 mothers who smoked during pregnancy with cutoff 0.3. CpG sites indices are given in Table 1.

in gene GFI1 and cg 04598670 (index number 4) in gene ENSG00000225718. We postulate that maternal smoking is likely to impose stronger dependence between CpG sites compared to subjects not exposed to maternal smoking. The size of the network for subjects exposed to maternal smoking is larger than that for nonexposed subjects. We do not expect this difference in network size is caused by sample size difference (197 vs. 48). Random samples of size 48 chosen from the 197 nonsmoking exposed girls resulted in networks with even smaller sizes, supporting the postulation of weaker dependence between CpG sites among nonsmokers compared to smokers.

5. Summary and discussion

In this article, we propose a Bayesian computational algorithm to infer directed links or networks among variables of interest, together with posterior probability. This algorithm is based on a more advanced sampler, Equi-Energy sampler. Moreover, graph orders, of which the space is much smaller than space of networks, are employed to build the Markov chain. The proposed algorithm has a better ability of escaping local traps and is expected to obtain more regular posterior distributions compared to existing FK algorithm.

It performs consistently well in terms of smaller number of false positives and false negatives, as seen in simulations. Furthermore, the proposed method has the potential to exclude noisy nodes which should not be in the network. This finding provides informative assistance in the interpretation of the estimated graphs using real data, in that some nodes are shown in the network for nonsmokers, but absent in the network for smokers. One potential limitation of the EE or SQEE is the computing efficiency. Although the proposed ASQEE improved the computing efficiency by use of one graph with a high probability, there is still a need to further reduce the computing time, and this is our ongoing work.

Acknowledgments

The project was supported by National Institutes of Health, NIH R01AI091905 (PI: W. Karmaus) and NIH R21AI099367 (PI: H. Zhang) for the work by H. Zhang and W. Karmaus, and the University of Memphis Center for Translational Informatics, FedEx Institute of Technology, and the Assisi Foundation of Memphis for the contribution of R Homayouni.

Appendix: Derivation of the marginal posterior of $\mathcal{G}$

The joint posterior of $\boldsymbol{\theta}^{\mathcal{G}}$, $\mathcal{G}$ is, up to a normalizing constant,

$$P(\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G}|\mathbb{X}) \propto P(\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G}) \times P(\mathbb{X}|\boldsymbol{\theta}^{\mathcal{G}}, \mathcal{G})$$

$$= \prod_{i=1}^{n} \gamma^{T_i} \left[\frac{\psi_i^{\delta_i}}{\Gamma(\delta_i)} \left(\sigma_i^2\right)^{-\delta_i - 1} e^{-\frac{\psi_i}{\sigma_i^2}} \left(2\pi\sigma_i^2\right)^{-\frac{n_i+1}{2}} \left|\left(\boldsymbol{X_i^{Pa}}^T \boldsymbol{X_i^{Pa}}\right)^{-1}\right|^{-\frac{1}{2}} \right.$$

$$\left. \times \exp\left\{ -\frac{\boldsymbol{\beta_i}^{\mathrm{T}} \boldsymbol{X_i^{Pa}}^{\mathrm{T}} \boldsymbol{X_i^{Pa}} \boldsymbol{\beta_i}}{2\sigma_{\mathrm{i}}^2} \right\} \prod_{\mathrm{h}=1}^{\mathrm{M}} \prod_{\mathrm{i}=1}^{\mathrm{n}} \left[\frac{1}{\sqrt{2\pi\sigma_{\mathrm{i}}^2}} \exp\left\{ -\frac{\left(\mathrm{x_{ih}} - \boldsymbol{\beta_i}^{\mathrm{T}} \boldsymbol{x_i^{Pa}}\right)^2}{2\sigma_{\mathrm{i}}^2} \right\} \right] \right.$$

$$\propto \prod_{i=1}^{n} \gamma^{T_i} \left[\left(\sigma_i^2\right)^{-\frac{M}{2}} \exp\left\{ -\frac{1}{2\sigma_{\mathrm{i}}^2} \left(\boldsymbol{X_i} - \boldsymbol{X_i^{Pa}} \boldsymbol{\beta_i}\right)^{\mathrm{T}} \left(\boldsymbol{X_i} - \boldsymbol{X_i^{Pa}} \boldsymbol{\beta_i}\right) \right\} \right.$$

$$\times \left(\sigma_i^2\right)^{-\frac{T_i+1}{2}} \exp\left\{-\frac{\boldsymbol{\beta_i}^{\mathrm{T}}(\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i^{Pa}})\boldsymbol{\beta_i}}{2\sigma_{\mathrm{i}}^2}\right\} \times \left(\sigma_{\mathrm{i}}^2\right)^{-\delta_{\mathrm{i}}-1} \exp\left\{-\frac{\psi_{\mathrm{i}}}{\sigma_{\mathrm{i}}^2}\right\}\Bigg]$$

$$= \prod_{i=1}^{n} \gamma^{T_i}\Bigg[\left(\sigma_i^2\right)^{-\frac{T_i+1}{2}} \exp\left\{-\frac{2\boldsymbol{\beta_i}^{\mathrm{T}}\left(\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i^{Pa}}\right)\boldsymbol{\beta_i} + \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{X_i} - \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{X_i^{Pa}}\boldsymbol{\beta_i} - \boldsymbol{\beta_i}^{\mathrm{T}}\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i}}{2\sigma_{\mathrm{i}}^2}\right\}$$

$$\times \left(\sigma_i^2\right)^{-(\frac{M}{2}+\delta_i)-1} \exp\left\{-\frac{\psi_{\mathrm{i}}}{\sigma_{\mathrm{i}}^2}\right\}\Bigg]$$

$$= \prod_{i=1}^{n} \gamma^{T_i}\Bigg[\left(\sigma_i^2\right)^{-\frac{T_i+1}{2}} \exp\left\{-\frac{\left(\boldsymbol{\beta_i} - \frac{1}{2}\boldsymbol{\eta_i}\right)^{\mathrm{T}}\left[2\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i^{Pa}}\right]\left(\boldsymbol{\beta_i} - \frac{1}{2}\eta_{\mathrm{i}}\right)}{2\sigma_{\mathrm{i}}^2}\right\}$$

$$\left(\sigma_i^2\right)^{-(\frac{M}{2}+\delta_i)-1} \exp\left\{-\frac{2\psi_{\mathrm{i}} - \frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i}\right)^{\mathrm{T}}\boldsymbol{\eta_i} + \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{X_i}}{2\sigma_{\mathrm{i}}^2}\right\}\Bigg] \tag{A.1}$$

where $\boldsymbol{\eta_i} = (\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}})^{-1}\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}$. The last equality is because

$$2\boldsymbol{\beta_i}^{T}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right)\boldsymbol{\beta_i} + \boldsymbol{X_i}^{T}\boldsymbol{X_i} - \boldsymbol{X_i}^{T}\boldsymbol{X_i^{Pa}}\boldsymbol{\beta_i} - \boldsymbol{\beta_i}^{T}\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}$$

$$= \left[\boldsymbol{\beta_i} - \frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right)^{-1}\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}\right]^{T}\left[2\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right]\left[\boldsymbol{\beta_i} - \frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right)^{-1}\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}\right]$$

$$-\frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}\right)^{T}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right)^{-1}\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i} + \boldsymbol{X_i}^{T}\boldsymbol{X_i}$$

$$= \left(\boldsymbol{\beta_i} - \frac{1}{2}\boldsymbol{\eta_i}\right)^{T}\left[2\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right]\left(\boldsymbol{\beta_i} - \frac{1}{2}\boldsymbol{\eta_i}\right) - \frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i}\right)^{T}\boldsymbol{\eta_i} + \boldsymbol{X_i}^{T}\boldsymbol{X_i}$$

By integrating out σ_i^2, $\boldsymbol{\beta_i}$, we get the marginal posterior distribution of $\mathcal{G}$,

$$P(\mathcal{G}|\mathbb{X}) \propto \prod_{i=1}^{n} \gamma^{T_i} \int_{\sigma_i^2}\int_{\boldsymbol{\beta_i}} \Bigg[\left(\sigma_i^2\right)^{-\frac{T_i+1}{2}} \exp\left\{-\frac{\left(\boldsymbol{\beta_i} - \frac{1}{2}\boldsymbol{\eta_i}\right)^{\mathrm{T}}\left[2\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i^{Pa}}\right]\left(\boldsymbol{\beta_i} - \frac{1}{2}\eta_{\mathrm{i}}\right)}{2\sigma_{\mathrm{i}}^2}\right\}$$

$$\times \left(\sigma_i^2\right)^{-(\frac{M}{2}+\delta_i)-1} \exp\left\{-\frac{2\psi_{\mathrm{i}} - \frac{1}{2}(\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i})^{\mathrm{T}}\boldsymbol{\eta_i} + \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{X_i}}{2\sigma_{\mathrm{i}}^2}\right\}\Bigg]$$

$$= \prod_{i=1}^{n} \gamma^{T_i}\Bigg[\pi^{\frac{T_i+1}{2}}\left|\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right|^{-\frac{1}{2}} \int_{\sigma_i^2} \left(\sigma_i^2\right)^{-(\frac{M}{2}+\delta_i)-1}$$

$$\times \exp\left\{-\frac{2\psi_{\mathrm{i}} - \frac{1}{2}\left(\boldsymbol{X_i^{Pa}}^{\mathrm{T}}\boldsymbol{X_i}\right)^{\mathrm{T}}\boldsymbol{\eta_i} + \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{X_i}}{2\sigma_{\mathrm{i}}^2}\right\}$$

$$= \prod_{i=1}^{n} \gamma^{T_i}\left[\pi^{\frac{T_i+1}{2}}\left|\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i^{Pa}}\right|^{-\frac{1}{2}} \times \frac{\Gamma\left(\frac{M}{2}+\delta_i\right)}{\left\{\psi_i - \frac{1}{4}(\boldsymbol{X_i^{Pa}}^{T}\boldsymbol{X_i})^{T}\boldsymbol{\eta_i} + \frac{1}{2}\boldsymbol{X_i}^{T}\boldsymbol{X_i}\right\}^{\frac{M}{2}+\delta_i}}\right] \square \tag{A.2}$$

References

Altomare, D., Consonni, G., La Rocca, L. (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics* 69:478–487.

Andersson, S. A., Madigan, D., Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* 25(2):505–541.

Arshad, S. H., Hide, D. W. (1992). Effect of environmental factors on the development of allergic disorders in infancy. *Journal of Allergy and Clinical Immunology* 90:235–241.

Balov, N. (2013). Consistent model selection of discrete Bayesian networks from incomplete data. *Electronic Journal of Statistics* 7:1047–1077.

Bouckaert, R. (1994). *Properties of Learning Algorithms for Bayesian Belief Networks*. The Netherlands: Utrecht University.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 2:445–498.

Chickering, D. M., Heckerman, D., Meek, C. (2004). Large-sample learning of Bayesian networks is np-hard. *Journal of Machine Learning Research* 5:1287–1330.

de Campos, P. C., Ji, Q. (2011). Efficient structure learning of Bayesian networks using constraints. *Journal of Machine Learning Research* 12:663–689.

Eaton, D. (2007). Bayesian structure learning using dynamic programming and MCMC. In Ron P., van der Gaag L., eds. *Proceeding UAI'07 Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. Vancouver, BC, Canada:AUAI Press Arlington, pp. 101–108.

Ellis, B., Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association* 103(482):778–789.

Friedman, N., Koller, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* 50:95–125.

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

Giudici, P., Green, P. J. (1999). Decomposable graphical gaussian model determination. *Biometrika* 86(4):785–801.

Grzegorczyk, M., Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* 71:265–305.

Han, S., Wong, R., Lee, T., Shen, L., Li, S.-Y., Fan, X. (2014). A full Bayesian approach for boolean genetic network inference. *PLoS ONE* 9(12):e115806.

Harris, N., Drton, M. (2013). Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research* 14:3365–3383.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109.

Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In: Jordan, M., ed. *Learning in Graphical Models*, Technical report, MIT Press.

Heckerman, D., Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243.

Joubert, B. R., Hberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., ivind Midttun Cupul-Uicab, L. A., Ueland, P. M., Wu, M. C., Nystad, W., Bell, D. A., Peddada, S. D., and London1, S. J. (2012). 450k epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives* 120:1425–1431.

Kalisch, M., Bhlmann, P. (2008). Robustification of the pc-algorithm for directed acyclic graphs. *Journal of Computational and Graphical Statistics* 17:773–789.

Kalisch, M., Bhlmann, P., Chickering, M. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* 8:613–636.

Kou, S., Zhou, Q., Wong, W. H. (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics* 34:1581–1619.

Larranaga, P., Poza, M., Yurramendi, Y., Murga, R. H., Kuijpers, C. M. H. (1996). Structure learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18:912–926.

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

Liang, F., Zhang, J. (2009). Learning Bayesian networks for discrete data. *Computational Statistics & Data Analysis* 53:865–876.

Madigan, D., Andersson, S., Perlman, M., Volinsky, C. (1996). Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. *Communications in Statistics: Theory and Methods*, 2493–2519.

Madigan, D., York, J., Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review* 63(2):215–232.

Moore, A., keen Wong, W. (2003). Optimal reinsertion: A new search operator for accelerated and more accurate Bayesian network structure learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML 03)*. Washington DC:ICML.

Philibert, R. A., Beach, S. R. H., Lei, M.-K., Brody, G. H. (2013). Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clinical Epigenetics* 5:19.

Shojaie, A., Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* 97:519–538.

Spirtes, P., Glymour, C., Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press.

Tsamardinos, I., Brown, L., Aliferis, C. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65:31–78.

Zhou, Q. (2011). Multi-domain sampling with applications to structural inference of Bayesian networks. *Journal of the American Statistical Association* 106:1317–1330.