

Learning Bayesian Networks for Biomedical Data

Faming Liang (*Texas A&M University*)

- Liang, F. and Zhang, J. (2009) Learning Bayesian Networks for Discrete Data. *Computational Statistics and Data Analysis*, **53**, 865-876.

The Bayesian network is a directed acyclic graph (DAG) in which the nodes represent the variables in the domain and the edges correspond to direct probabilistic dependencies between them.

Researchers have directed interest in Bayesian networks and applications of such models to biological data, see e.g., Friedman *et al.* (2000) and Ellis and Wong (2008).

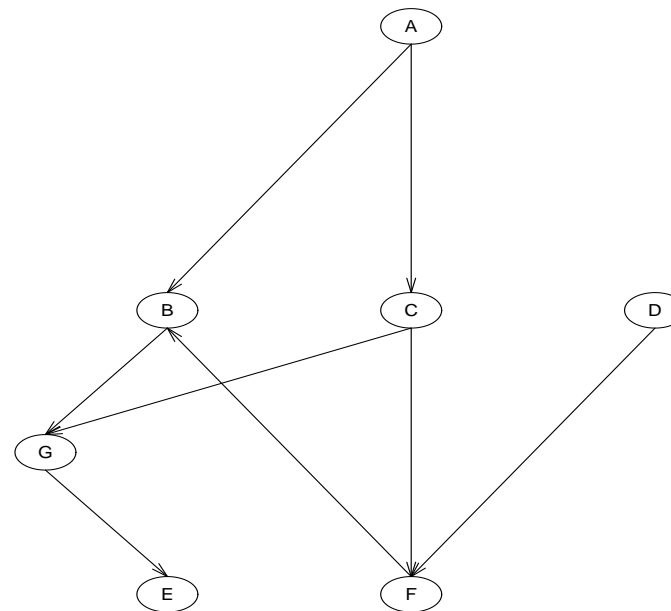


Figure 1: An example of Bayesian networks: the network is compatible with both the orders ACDFBGE and ADCFBGE.

Bayesian Network Learning Approaches:

- **Conditional independence test-based approach:** The networks constructed by the approach are usually asymptotically correct, but as pointed out by Cooper and Herskovits (1992) that the conditional independence tests with large condition-sets may be unreliable unless the volume of data is enormous.
- **Optimization-based approach:** It attempts to find a network that optimizes a selected scoring function, such as entropy, minimum description length, and Bayesian scores. The optimization procedures employed are usually heuristic, which often stops at a local optimal structure.

- **MCMC simulation-based approach:**

- The Metropolis-Hastings simulation based approach (Madigan and Raftery, 1994): it tends to be trapped by a local optimal structure.
- Two stage approach (Friedman and Koller, 2003): use the MH algorithm to sample a temporal order of nodes, and then sample a network structure compatible with the given node order.

The structures sampled does not follow the correct posterior distribution, as the temporal order does not induce a partition of the space of network structures (Ellis and Wong, 2008). A network may be compatible with more than one order.

Temporal order: for any Bayesian network, there exists a temporal order of nodes such that for any two nodes X and Y , if there is an edge from X and Y , then X must be preceding to Y in the order.

A Bayesian network model can be defined as a pair $B = (\mathcal{G}, \rho)$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed acyclic graph that represents the structure of the network, \mathcal{V} denotes the set of nodes, \mathcal{E} denotes the set of edges, and ρ is a vector of conditional probabilities.

For a node V , a parent of V is a node from which there is a directed link to V . The set of parents of V is denoted by $pa(V)$. The joint distribution of the variables $\mathbf{V} = \{V_1, \dots, V_d\}$ can be specified by the decomposition

$$P(\mathbf{V}) = \prod_{i=1}^d P(V_i | pa(V_i)). \quad (1)$$

- Let $\mathcal{D} = \{\mathbf{V}_1, \dots, \mathbf{V}_N\}$ denote a set of independently and identically distributed samples drawn from the distribution (1).
- Let n_{ijk} count the number of samples for which V_i is in state j and $pa(V_i)$ is in state k . Then

$$(n_{i1k}, \dots, n_{ir_ik}) \sim \text{Multinomial}\left(\sum_{j=1}^{r_i} n_{ijk}, \rho_{ik}\right), \quad (2)$$

where $\rho_{ik} = (\rho_{i1k}, \dots, \rho_{ir_ik})$, and ρ_{ijk} is the probability of variable V_i in state j conditioned on that $pa(V_i)$ is in state k .

- The likelihood function of the Bayesian network model is given by

$$P(\mathcal{D}|\mathcal{G}, \rho) = \prod_{i=1}^d \prod_{k=1}^{q_i} \binom{\sum_{j=1}^{r_i} n_{ijk}}{n_{i1k}, \dots, n_{ir_ik}} \rho_{i1k}^{n_{i1k}} \dots \rho_{ir_ik}^{n_{ir_ik}}. \quad (3)$$

- Since a network with a large number of edges is often less interpretable and there is a risk of over-fitting, it is important to use priors over the network space that encourages sparsity.

$$P(\mathcal{G}|\beta) \propto \left(\frac{\beta}{1-\beta}\right)^{\sum_{i=1}^d |pa(V_i)|}, \quad (4)$$

where $0 < \beta < 1$ is a user-specified parameter. We set $\beta = 0.1$ for all examples.

- The parameters ρ is subject to a product Dirichlet distribution

$$P(\rho|\mathcal{G}) = \prod_{i=1}^d \prod_{k=1}^{q_i} \frac{\Gamma(\sum_{j=1}^{q_i} \alpha_{ijk})}{\Gamma(\alpha_{i1k}) \cdots \Gamma(\alpha_{ir_i k})} \rho_{i1k}^{\alpha_{i1k}-1} \cdots \rho_{ir_i k}^{\alpha_{ir_i k}-1}, \quad (5)$$

where $\alpha_{ijk} = 1/(r_i q_i)$ as used by Ellis and Wong (2008).

- Combining with the likelihood function and the prior, we get the posterior

$$P(\mathcal{G}|\mathcal{D}) \propto \prod_{i=1}^d \left(\frac{\beta}{1-\beta}\right)^{|pa(V_i)|} \prod_{k=1}^{q_i} \frac{\Gamma(\sum_{j=1}^{r_i} \alpha_{ijk})}{\Gamma(\sum_{j=1}^{r_i} (\alpha_{ijk} + n_{ijk}))} \prod_{j=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})},$$

(6)

Bayesian network and causal Bayesian network

The Bayesian network is conceptually different from the causal Bayesian network. In the causal Bayesian network, each edge can be interpreted as a direct causal relation between a parent node and a child node, relative to the other nodes in the network (Pearl, 1988). The formulation of Bayesian networks, as described above, is not sufficient for causal inference. To learn a causal Bayesian network, one needs a dataset obtained through experimental interventions. In general, one cannot learn a causal Bayesian network from the observational data alone.

Algorithm Setting

- Suppose that we are working with the following Boltzmann distribution,

$$f(x) = \frac{1}{Z} \exp \{ -U(x)/\tau \}, \quad x \in \mathcal{X}, \quad (7)$$

where Z is the normalizing constant, τ is the temperature, \mathcal{X} is the sample space, and $U(x) = -\log P(\mathcal{G}|\mathcal{D})$ for Bayesian networks. For Bayesian networks, the sample space \mathcal{X} is finite.

- Let $\psi(x) = \exp(-U(x)/\tau)$ be a non-negative function defined on the sample space with $0 < \int_{\mathcal{X}} \psi(x) dx < \infty$, and $\theta_i = \log(\int_{E_i} \psi(x) dx)$.

- Suppose that the sample space has been partitioned into m disjoint subregions denoted by $E_1 = \{x : U(x) \leq u_1\}$, $E_2 = \{x : u_1 < U(x) \leq u_2\}$, \dots , $E_{m-1} = \{x : u_{m-2} < U(x) \leq u_{m-1}\}$, and $E_m = \{x : U(x) > u_{m-1}\}$, where u_1, \dots, u_{m-1} are real numbers specified by the user.
- Let $\pi = (\pi_1, \dots, \pi_m)$ be an m -vector with $0 < \pi_i < 1$ and $\sum_{i=1}^m \pi_i = 1$, which will be called the desired sampling distribution.

- Let $\{\gamma_t\}$ be a positive, non-decreasing sequence satisfying the conditions,

$$(i) \sum_{t=0}^{\infty} \gamma_t = \infty, \quad (ii) \sum_{t=0}^{\infty} \gamma_t^{\delta} < \infty, \quad (8)$$

for some $\delta \in (1, 2)$. In our simulations, we set

$$\gamma_t = \frac{t_0}{\max(t_0, t)}, \quad t = 0, 1, 2, \dots \quad (9)$$

- Let $\{\mathcal{K}_s, s \geq 0\}$ be a sequence of compact subsets of Θ such that

$$\bigcup_{s \geq 0} \mathcal{K}_s = \Theta, \quad \text{and} \quad \mathcal{K}_s \subset \text{int}(\mathcal{K}_{s+1}), \quad s \geq 0, \quad (10)$$

where $\text{int}(A)$ denotes the interior of set A .

- Let \mathcal{X}_0 be a subset of \mathcal{X} , and let $\mathcal{T} : \mathcal{X} \times \Theta \rightarrow \mathcal{X}_0 \times \mathcal{K}_0$ be a measurable function which maps a point in $\mathcal{X} \times \Theta$ to a random point in $\mathcal{X}_0 \times \mathcal{K}_0$.
- Let σ_k denote the number of truncations performed until iteration k . Let \mathcal{S} denote the collection of the indices of the subregions from which a sample has been proposed; that is, \mathcal{S} contains the indices of all subregions which are known to be non-empty.

The SAMC algorithm

- (a) (Sampling) Simulate a sample $x^{(t+1)}$ by a single MH update with the target distribution

$$f_{\theta^{(t)}}(x) \propto \sum_{i=1}^{m-1} \frac{\psi(x)}{e^{\theta_i^{(t)}}} I(x \in E_i) + \psi(x) I(x \in E_m), \quad (11)$$

where $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_{m-1}^{(t)})$.

- (a.1) Generate y according to a proposal distribution $q(x_t, y)$. If $J(y) \notin \mathcal{S}$, set $\mathcal{S} \leftarrow \mathcal{S} + \{J(y)\}$.
- (a.2) Calculate the ratio

$$r = e^{\theta_{J(x^{(t)})}^{(t)} - \theta_{J(y)}^{(t)}} \frac{\psi(y)q(y, x^{(t)})}{\psi(x^{(t)})q(x^{(t)}, y)}.$$

- (a.3) Accept the proposal with probability $\min(1, r)$. If it is accepted, set $x^{(t+1)} = y$; otherwise, set $x^{(t+1)} = x^{(t)}$.

(b) (Weight updating) For all $i \in \mathcal{S}$, set

$$\theta_i^{(t+\frac{1}{2})} = \theta_i^{(t)} + a_{t+1} \left(I_{\{x^{(t+1)} \in E_i\}} - \pi_i \right) - a_{t+1} \left(I_{\{x^{(t+1)} \in E_m\}} - \pi_m \right). \quad (12)$$

(c) (Varying truncation) If $\theta^{(t+\frac{1}{2})} \in \mathcal{K}_{\sigma_t}$, then set $(\theta^{(t+1)}, x^{(t+1)}) = (\theta^{(t+\frac{1}{2})}, x^{(t+1)})$ and $\sigma_{t+1} = \sigma_t$; otherwise, set $(\theta^{(t+1)}, x^{(t+1)}) = \mathcal{T}(\theta^{(t)}, x^{(t)})$ and $\sigma_{t+1} = \sigma_t + 1$.

Self-adjusting mechanism: If a proposal is rejected, the log-weight of the subregion that the current sample belongs to will be adjusted to a larger value, and thus the proposal of jumping out from the current subregion will less likely be rejected in the next iteration.

This mechanism enables SAMC to escape from local energy minima very quickly.

Proposal Distribution:

The proposal distribution $q(x, y)$ used in the MH updates satisfies the following condition: For every $x \in \mathcal{X}$, there exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that

$$|x - y| \leq \epsilon_1 \implies q(x, y) \geq \epsilon_2, \quad (13)$$

where $|x - y|$ denotes a certain distance measure between x and y .

This is a natural condition in study of MCMC theory (Roberts and Tweedie, 1996).

Convergence:

Under the conditions (8) and (13), for all non-empty subregions,

$$\theta_i^{(t)} \rightarrow C + \log \left(\int_{E_i} \psi(x) dx \right) - \log (\pi_i + \pi_0), \quad (14)$$

as $t \rightarrow \infty$, where $\pi_0 = \sum_{j \in \{i: E_i = \emptyset\}} \pi_j / (m - m_0)$, $m_0 = \#\{i : E_i = \emptyset\}$ is the number of empty subregions, and $C = -\log \left(\int_{E_m} \psi(x) dx \right) + \log (\pi_m + \pi_0)$.

Estimation:

Let $(x^{(1)}, \theta^{(1)}), \dots, (x^{(n)}, \theta^{(n)})$ denote a set of samples generated by SAMC. For an integrable function $h(x)$, the expectation $E_f h(x)$ can be estimated by

$$\widehat{E_f h(x)} = \frac{\sum_{t=1}^n e^{\theta^{(t)} J(x^{(t)})} h(x^{(t)})}{\sum_{t=1}^n e^{\theta^{(t)} J(x^{(t)})}}. \quad (15)$$

As $n \rightarrow \infty$, $\widehat{E_f h(x)} \rightarrow E_f h(x)$ for the same reason that the usual importance sampling estimate converges.

Let \mathcal{G} denote a feasible Bayesian network for the data \mathcal{D} . At each iteration of SAMC, the sampling step can be performed as follows:

- (a) Uniformly randomly choose between the following possible changes to the current network $\mathcal{G}^{(t)}$ producing \mathcal{G}' :
 - (a.1) Temporal order change: Swap the order of two neighboring models. If there is an edge between them, reverse its direction.
 - (a.2) Skeletal change: Add (or delete) an edge between a pair of randomly selected nodes.
 - (a.3) Double skeletal change: Randomly choose two different pairs of nodes, and add (or delete) edges between each pair of the nodes.

(b) Calculate the ratio

$$r = e^{\theta_{J(\mathcal{G}')}^{(t)} - \theta_{J(\mathcal{G}^{(t)})}^{(t)}} \frac{\psi(\mathcal{G}')}{\psi(\mathcal{G}^{(t)})} \frac{T(\mathcal{G}' \rightarrow T(\mathcal{G}^{(t)}))}{T(\mathcal{G}^{(t)} \rightarrow \mathcal{G}')},$$

where $\psi(\mathcal{G})$ is defined as the right hand side of (6), and the ratio of the proposal probabilities $T(\mathcal{G}' \rightarrow T(\mathcal{G}^{(t)})) / T(\mathcal{G}^{(t)} \rightarrow \mathcal{G}') = 1$ for all of the three types of the changes. Accept the new network structure \mathcal{G}' with probability $\min(1, r)$. If it is accepted, set $\mathcal{G}^{(t+1)} = \mathcal{G}'$; otherwise, set $\mathcal{G}^{(t+1)} = \mathcal{G}^{(t)}$.

Let $h(\mathcal{G})$ denote a quantity of interest for a Bayesian network, such as the presence/absence of an edge or a future observation.

The expectation of $h(\mathcal{G})$ with respect to the posterior (6) can be estimated by

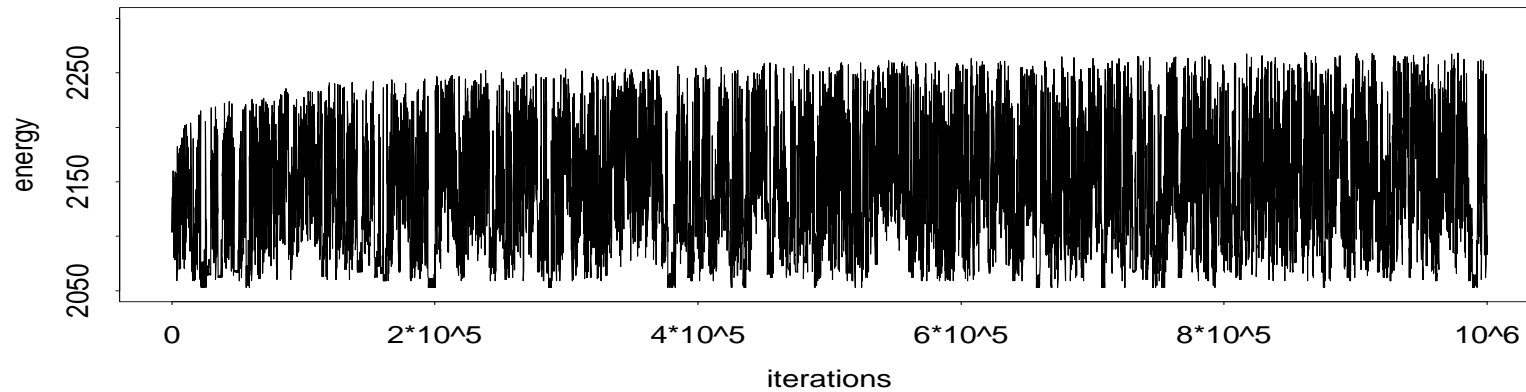
$$\widehat{E_P h(\mathcal{G})} = \frac{\sum_{k=n_0+1}^n h(\mathcal{G}_k) e^{\theta_{J(\mathcal{G}_k)}^{(k)}}}{\sum_{k=n_0+1}^n e^{\theta_{J(\mathcal{G}_k)}^{(k)}}}, \quad (16)$$

where $(\mathcal{G}_{n_0+1}, \theta_{J(\mathcal{G}_{n_0+1})}^{(n_0+1)}), \dots, (\mathcal{G}_n, \theta_{J(\mathcal{G}_n)}^{(n)})$ denotes a set of samples generated by SAMC, and n_0 denotes the number of burn-in iterations.

Simulated example

Suppose that a dataset, consisting of 500 independent observations, has been generated from the network (shown in Figure 1) according to certain distributions.

(a) Evolving path of SAMC samples



(b) Evolving path of MH samples

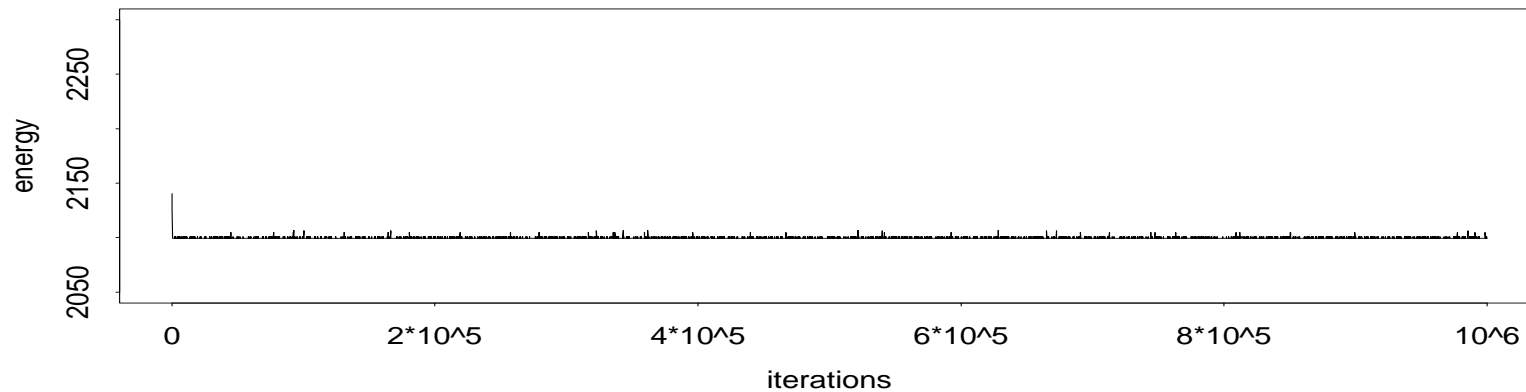


Figure 2: The sample paths (in the space of energy) produced by SAMC (upper panel) and MH (lower panel) for the simulated example.

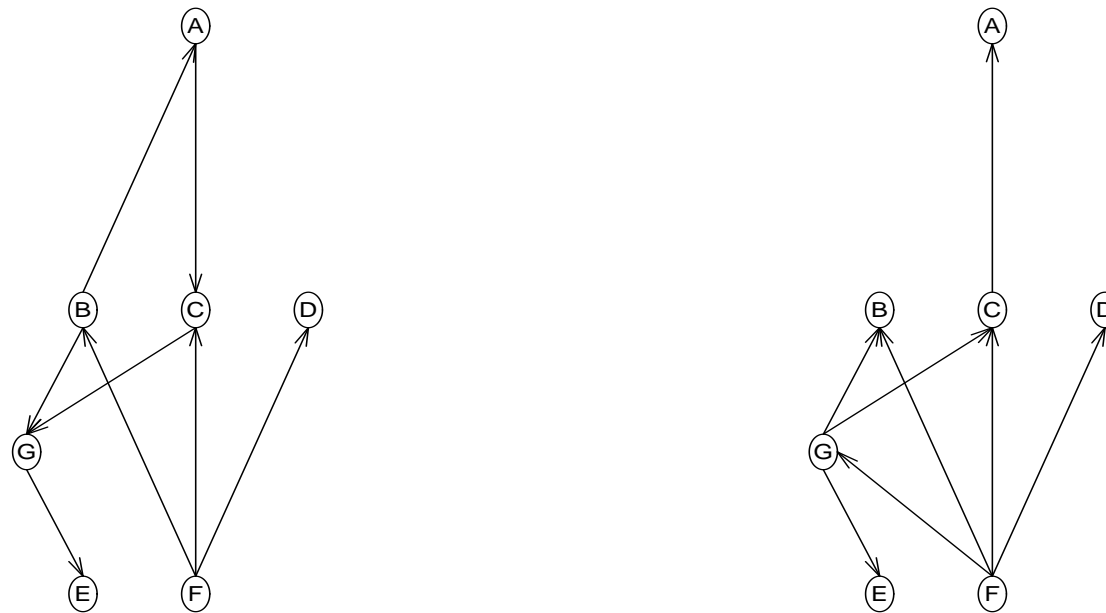


Figure 3: The highest posteriori network structures produced by SAMC (left panel) and MH (right panel) for the simulated example.

Table 1: Estimates of edge presence probabilities for the network in Fig 1.

	A	B	C	D	E	F	G
A	—	0 (0)	0.9997 (0.0001)	0 (0)	0 (0)	0 (0)	0 (0)
B	1 (0)	—	0 (0)	0 (0)	0.0046 (0.0009)	0.4313 (0.0552)	1 (0)
C	0.0003 (0.0001)	0 (0)	—	0 (0)	0 (0)	0 (0)	0.9843 (0.0036)
D	0 (0)	0 (0)	0 (0)	—	0.0002 (0)	0.0476 (0.0233)	0 (0)
E	0 (0)	0 (0)	0 (0)	0 (0)	—	0 (0)	0.0044 (0.0009)
F	0 (0)	0.5687 (0.0552)	1 (0)	0.9524 (0.0233)	0.1638 (0.0184)	—	0 (0)
G	0 (0)	0 (0)	0.0003 (0.0001)	0 (0)	0.9956 (0.0009)	0 (0)	—

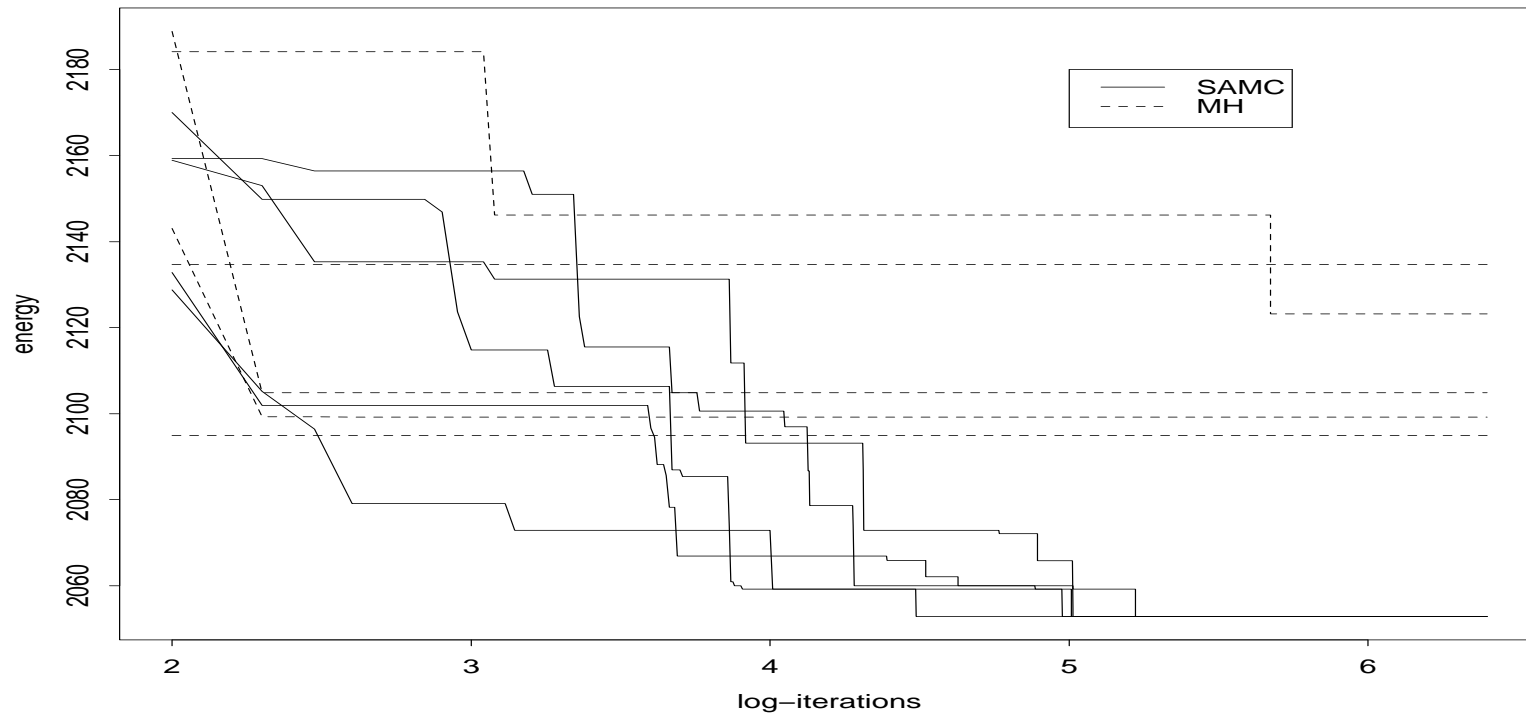


Figure 4: Progression paths of minimum energy values produced in the five runs of SAMC (solid lines) and in the five runs of MH (dashed lines) for the simulated example.

Wisconsin Breast Cancer Data

The Wisconsin Breast Cancer dataset has 683 samples, which consist of visually assessed nuclear features of fine needle aspirates taken from patients' breasts. Each sample was assigned a 9-dimensional vector of diagnostic characteristics: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Each component is in the interval 1 to 10, with 1 corresponding to the normal state and 10 to the most abnormal state. The samples were classified into two classes, benign and malignant.

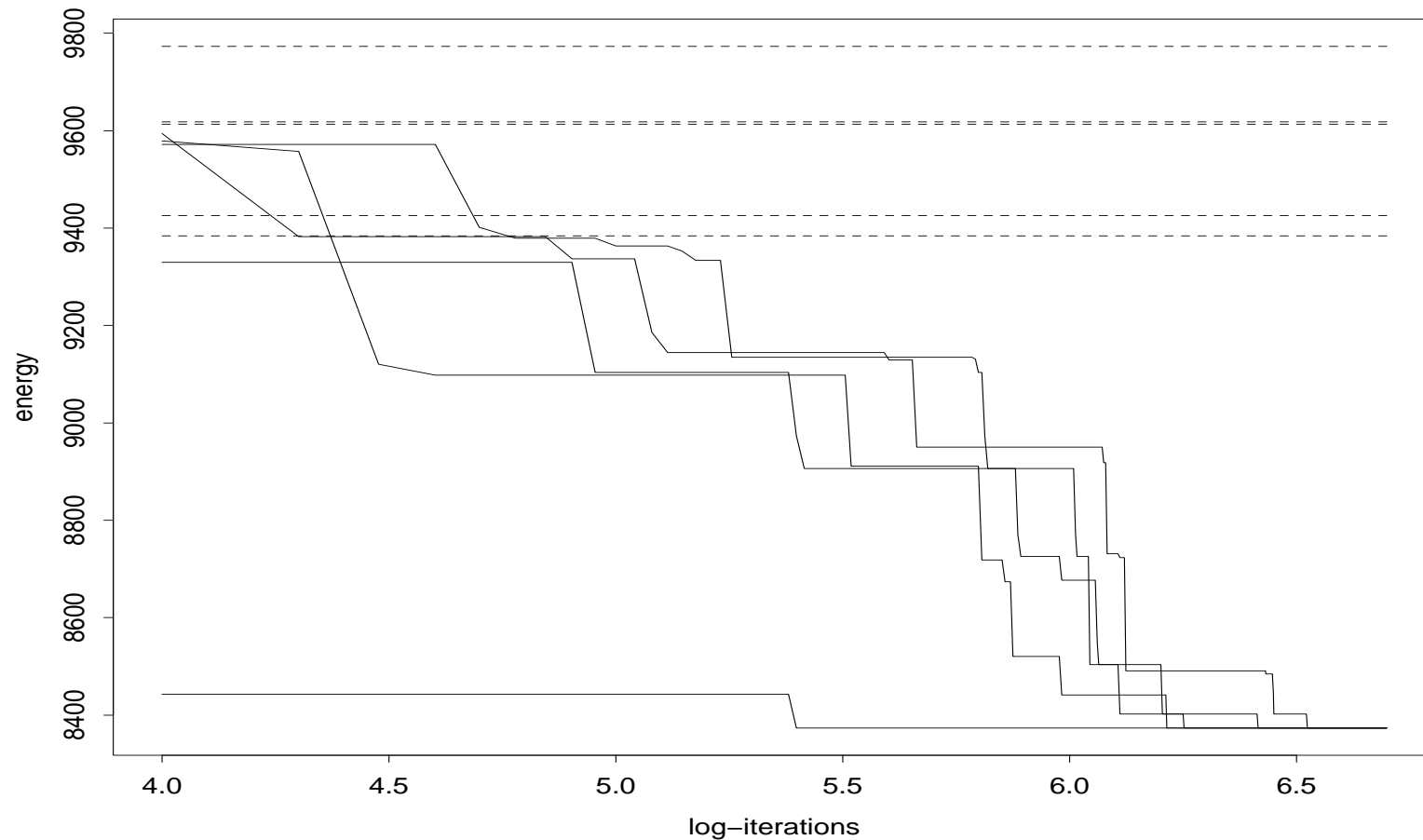


Figure 5: Progression paths of minimum energy values produced in five runs of SAMC (solid lines) and five runs of MH (dashed lines) for the Wisconsin Breast Cancer example.

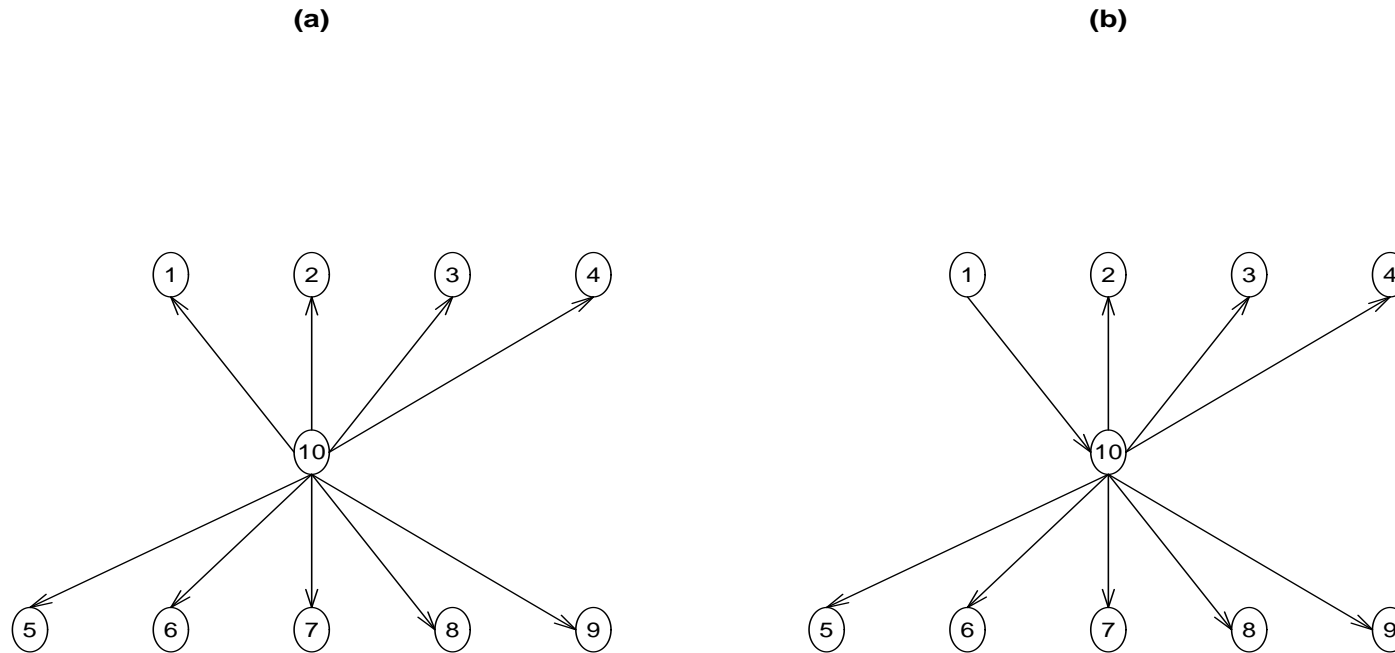


Figure 6: The putative MAP Bayesian network (left panel, with an energy value of 8373.9) and a suboptimal Bayesian network (right panel, with an energy value of 8441.73) produced by SAMC for the Wisconsin Breast Cancer data.

SPECT Heart Data

This dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 22 binary feature patterns were created for each patient.

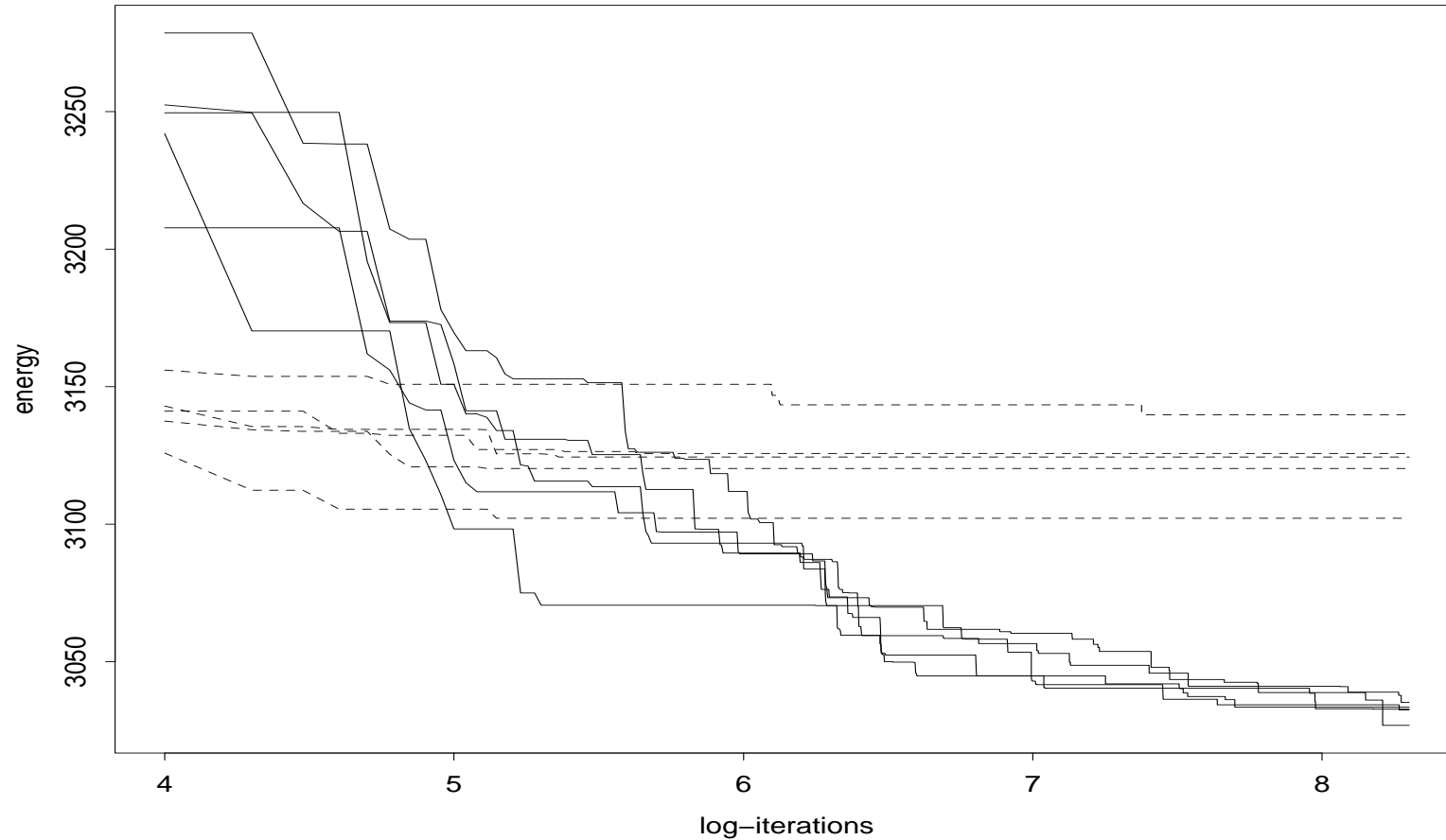


Figure 7: Progression paths of minimum energy values produced in five runs of SAMC (solid lines) and five runs of MH (dashed lines) for the SPECT data.

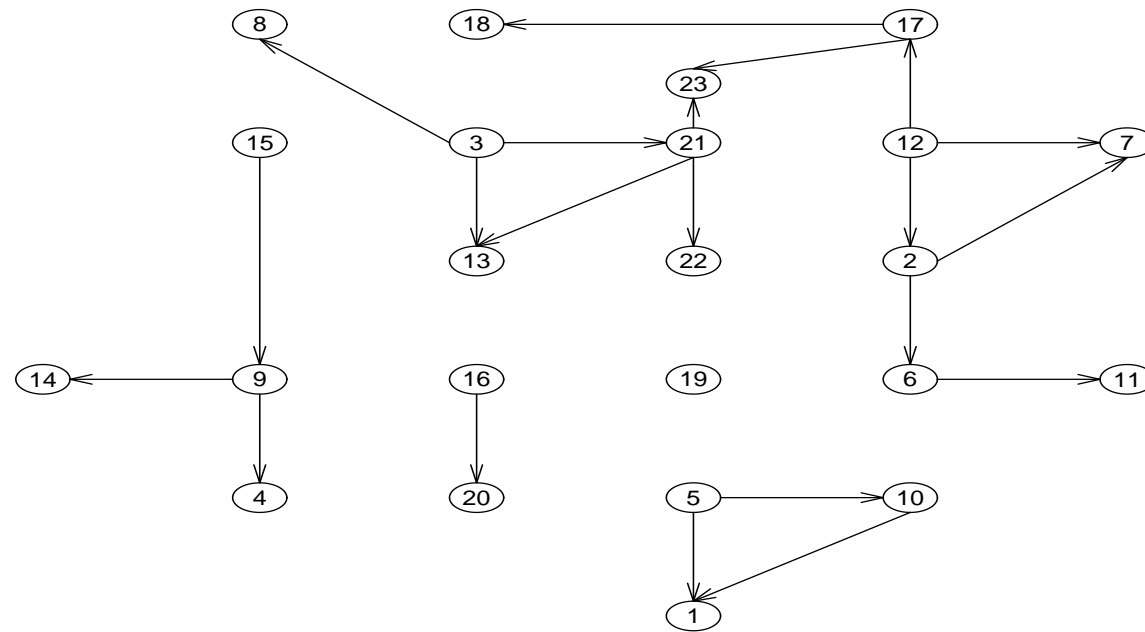


Figure 9: The consensus Bayesian network learned by SAMC for the SPECT data.

- The method can be extended to continuous data by appropriate discretization. Application of Bayesian networks to continuous data will be a future research direction.
- The computation can be accelerated by specifying a uniform desired sampling distribution. In this case, the weight updating step can be replaced by

$$\theta_i^{(t+\frac{1}{2})} = \theta_i^{(t)} + a_{t+1} \left(I_{\{x^{(t+1)} \in E_i\}} - I_{\{x^{(t+1)} \in E_m\}} \right), \quad (17)$$

where the weight is only updated for a single subregion that the current sample $x^{(t+1)}$ belongs to instead of all nonempty subregions.