Published in final edited form as:

Stat Probab Lett. 2011 February 1; 81(2): 220–230. doi:10.1016/j.spl.2010.11.009.

A Gaussian Mixed Model for Learning Discrete Bayesian Networks

Nikolay Balov^a

^a Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Ave, Rochester, NY-14642

Abstract

In this paper we address the problem of learning discrete Bayesian networks from noisy data. Considered is a graphical model based on mixture of Gaussian distributions with categorical mixing structure coming from a discrete Bayesian network. The network learning is formulated as a Maximum Likelihood estimation problem and performed by employing an EM algorithm. The proposed approach is relevant to a variety of statistical problems for which Bayesian network models are suitable - from simple regression analysis to learning gene/protein regulatory networks from microarray data.

Keywords

Bayesian networks; learning; mixture models; MLE; EM algorithm

1. Introduction

Graphical models provide a framework for representing the dependence between random variables through directed edges in graphs. Bayesian networks form a class of graphical models subject to the assumption that the variables can be ordered according to their causal relationship. They witnessed a renewed interest in recent years due to promising applications to systems biology and genetics - for example, Bayesian networks provide efficient means for describing the mechanisms of control and regulation of gene expressions.

Over the last few decades, a large amount of research and literature has been devoted to the problem of learning Bayesian networks - some main sources are Buntine (1996), Heckerman et al. (1995) and Neapolitan (2003). A typical learning is based on the so called score-based approach. It identifies a network structure according to some scoring criterion such as the *a posteriori* probability of a network given some data, or some likelihood based criteria such as AIC or BIC. Another learning approach is formed by the constraint-based algorithms that use conditional independence tests to find associations between the node-variables. Once the graphical structure of a Bayesian network is identified, in the context of a specified probability model, it is usually straightforward to estimate the parameters of the corresponding conditional probabilities. However, finding the structure of Bayesian

---

Correspondence to: Nikolay Balov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

networks with more than one parent is a NP-hard problem, Chickering (1996), which motivates the search of model constraints and assumptions to relax its general intractability.

In this paper we restrict our attention to discrete Bayesian networks with multinomial conditional distributions. The data is assumed to be observations on categorical variables plus additional Gaussian noise distributed independently among them. In other words, the nodes of the networks are assumed latent, unobserved, random variables, while the actual, observed, variables are their noisy instances. In this case, the conditional distributions of the observed continuous variables are essentially mixture of Gaussian ones and the model can be referred to as mixed Bayesian network. The model estimation is performed in a frequentist manner using the Maximum Likelihood (ML) principle. Under the assumption that the causal order of the node-variables is known, the MLE solution is approximated by employing an Expectation Maximization (EM) algorithm. The node order assumption, despite being considered unacceptable from a general learning standpoint, is the one that allows the global MLE, if exists, to be found, in contrast to many, more general but heuristic learning algorithms (see Heckerman et al. (1995) for systematical overview), which provide only 'local' solutions.

The model considered in this paper can be formulated as a special case of the Conditional Gaussian (CG) model, Lauritzen (1992), not to be mistaken with the substantially different linear Gaussian model. CG models belongs to the family of mixed graphical association models that have been first introduced by Lauritzen & Wermuth (1984). A CG model allows both discrete and continuous variables, but the latter can have as parents only discrete ones on which they depend through conditional Gaussian distributions. CG models are well studied and efficient local procedures for calculating their marginal distributions, means and variances are available, Lauritzen (1992) and Lauritzen & Jensen (2001). However, a method for estimating both the graphical and probability structures of a general CG model is not available.

In the context of network learning, the EM algorithm is typically applied in two settings - when some of the data is missing, Lauritzen (1995), or when some of the variables are hidden. In Chickering & Heckerman (1996), considered is a general approach to parameter estimation for networks with hidden variables that resorts to computationally intensive methods, such as Monte-Carlo, for approximation of the marginal distributions. The lack of close form expression of the marginals in a Bayesian network model limits substantially the applicability of EM.

The application of EM has been mostly limited to learning the probability model parameters of Bayesian networks. One exception is Friedman (1997), where a method for learning the structure of Discrete Bayesian networks from incomplete data is proposed. The algorithm described there is a variant of the EM in which the M-step is replaced by two alternating steps: a local structure search, similar to the greedy hill-climbing procedure, and parameter estimation. Its advantage is that it does not require the node order to be known. However, since it converges to locally optimal solutions, it is sensitive to the initial configurations, and second, requires discrete data. In contrast, the proposed here estimation procedure seamlessly combines the discretization and structure learning steps in one unified EM algorithm.

The main advantage of the framework considered in this paper is its ability to model continuous variables by unrestricted categorical distributions and thus able to express complex, non-linear relationships between them (see Example 1 below). Consequently, in appropriate settings, this approach provides greater representation power than the discrete Bayesian networks alone and parametric models such as the linear Gaussian networks.

Moreover, with close form expression for the expected sufficient statistics, our model estimation algorithm can be implemented accurately and efficiently.

The paper is organized as follows. In Section 2 we introduce the discrete Bayesian network model and in Section 3 its extension - the mixture of Gaussian Bayesian network (MGBN). In Section 4 we formulate the estimation problem for MGBN and derive an EM algorithm to solve it. Finally, in Section 5, we present simulation results and a small study for learning the interactions between some genes complicated in lung cancer.

The MGBN model inference is implemented in the *mugnet* R-package, which also contains code for reproducing the presented in the paper examples.

Example 1. (Motivation Example)

The MGBN is designed to model random vectors, which components are multi-modal continuous variables that follow some mixture of Gaussian distributions. In Table 1 we show the distribution of a random vector (X, Y) , such that $X = A + e_A$ and $Y = B + e_B$ for independent $e_A \sim \mathcal{N}(0, 0.04)$ and $e_B \sim \mathcal{N}(0, 0.08)$, and discrete random variables A and B with probability table fully specified by $P(A)$ and $P(B|A)$, shown on the left. On the right, plotted is a large sample of (X, Y) , which clearly shows the cluster-like structure of the joint distribution. Note that in this case fitting the linear Gaussian model $B \sim A$ is not appropriate. In the remainder of this paper, we present a method for recovering not only the distribution of (A, B) from a (X, Y) -sample, but also the dependence relations in case of many node-variables.

2. Discrete Bayesian Networks

A Bayesian Network (BN) $\mathcal{G} = (G, P)$ is defined by directed acyclic graph (DAG) G with random node-variables $x_i, i = 1, \dots, n$, and probability model P . The directed arcs of G determine the parent-child relations, while the model P is a set of conditional distributions, one for each node, specifying the distribution of the nodes conditional on their parents. $\mathcal{G}$ is discrete BN if each node x is a discrete random variable that takes values in some finite set of categories C_x and the conditional distributions are categorical. Let us denote with Π_i the set of parent indices for node x_i and with $P(x_i|x_j, j \in \Pi_i)$, or more concise $P(x_i|\Pi_i)$, its conditional probability. There is a topological order of the nodes of $\mathcal{G}$ such that the parents of each one appear always earlier in that order. To simplify the analysis, throughout this paper we will assume, without loss of generality, that the topological order of the nodes in G is the same as their index order, a fact expressed symbolically as $x_1 < x_2 < \dots < x_n$. That is, for every $i > 1$, $\Pi_i \subset \{1, 2, \dots, i-1\}$ and $\Pi_1 = \emptyset$. An application of the Bayes formula gives us the following factorization of the joint probability

$$\log P(x_1, \dots, x_n | \mathcal{G}) = \sum_{i=1}^n \log P(x_i | x_j, j < i) = \sum_{i=1}^n \log P(x_i | \Pi_i). \quad (1)$$

Hereafter, we assume that the node categorical sets are $C_i = \{1, \dots, a_i\}$ for some natural numbers $a_i > 1$. We also denote by $C[\Pi_i]$ the set of possible combination of parent categories, $C[\Pi_i] \equiv \otimes_{j \in \Pi_i} C_j$.

Categorical distributions, also called multinomial distributions, are the most general model for the conditional probabilities $P(x_i|\Pi_i)$. With a slight abuse of notation, with $P(x_i|\Pi_i = C)$ we will denote $P(x_i|x_j = c_j, j \in \Pi_i)$, for a parent configuration $C = (c_j) \in C[\Pi_i]$. By definition, $\sum_{c \in C_i} P(x_i = c | \Pi_i = C) = 1$. Hence, each conditional probability $P(x_i|\Pi_i)$ requires

$\alpha_i = (\alpha_i - 1) \times \prod_{j \in \Pi_i} \alpha_j$ parameters to be specified. The total number $\alpha = \sum_{i=1}^n \alpha_i$, called network complexity, gives the overall number of parameters of the discrete BN $\mathcal{G}$.

As defined, the discrete BNs have no parametric constraints on their conditional distributions except the number of node categories. However, this generality is counterbalanced by the potential information loss in the process of data categorization. In practice, most data is obtained by analogous instruments and is measured effectively in continuous form. In order to fit a discrete BN one needs, before anything else, to discretize the continuous outcome, which usually results in information loss and hampers the statistical analysis. This loss can be especially costly when dealing with small sample sizes, as it is the case with microarray data. In the existing practice, the problems of discretization and fitting a discrete BN, are considered separately, which is not a statistically satisfactory approach. The methodology we present next addresses this issue.

3. Mixture of Gaussian Bayesian Networks

We consider the following model

$$y_i = \beta_{i,x_i} + e_i, \quad i = 1, \dots, n, \quad (2)$$

where $y = (y_1, \dots, y_n)$ is a continuous random vector, $x = (x_1, \dots, x_n)$ comes from a discrete BN $\mathcal{G}$, and e_i are independent, zero mean Gaussian random variables with variances σ_i^2 , i.e. $e = (e_1, \dots, e_n) \sim \mathcal{N}(0, \text{diag}(\sigma_i^2))$. In this setting, y is the observed random vector, while x is unobserved, also latent, discrete random vector. The triple

$$\theta = (\{\beta_i = (\beta_{i,k})_{k \in C_i}\}_{i=1}^n, \{\sigma_i^2\}_{i=1}^n, \mathcal{G})$$

constitutes the parameter set for model (2).

For given x_i , y_i is a Gaussian variable with mean β_{i,x_i} and variance σ_i^2 . In equivalent formulation of (2), the logarithm of the density of y_i conditional on x_i can be written as

$$\log(P(y_i|x_i, \theta)) = -\frac{1}{2} [\log(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2} (y_i - \beta_{i,x_i})^2]. \quad (3)$$

The local Markov property implies that each x_i depends on $x_{j < i}$ (all x_j with $j < i$), only through $x_j \in \Pi_i$ (all x_j with $j \in \Pi_i$). Moreover, y_i depends on $x_{j \leq i}$ only through x_i . By applying these observations we obtain a convenient expression for the logarithm of the joint distribution of x and y

$$\log P(y, x|0) = \log P(y_1, x_1|\theta) + \sum_{i=2}^n \log P(y_i, x_i | y_{j < i}, x_{j < i}, \theta) = \sum_{i=1}^n [\log P(y_i|x_i, \theta) + \log P(x_i|x_{j \in \Pi_i}, \theta)] = \sum_{i=1}^n \log P(y_i, x_i | x_{j \in \Pi_i}, \theta). \quad (4)$$

By combining equations (1), (3) and (4), we write the log-likelihood of θ with respect to a given sample (y, x) as

$$l(\Theta|y, x) = -\frac{1}{2} \sum_{i=1}^n \left[\log(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2} (y_i - \beta_{i,x_i})^2 \right] + \log P(x_1, \dots, x_n | \mathcal{G}). \quad (5)$$

Some of the properties of model (2) can be elucidated by deriving the conditional distributions of y 's. For example, the distribution of y_i conditional on its parents $y_{j \in \Pi_i}$ is given by

$$E_{x_i|y_{j \in \Pi_i}} P(y_i|x_i, y_{j \in \Pi_i}) = E_{x_i|y_{j \in \Pi_i}} P(y_i|x_i) = \sum_{c \in C_i} P(y_i|x_i=c) P(x_i=c|y_{j \in \Pi_i}),$$

where $P(y_i|x_i)$ are Gaussian distributions, and thus, is a mixture of Gaussian ones. In the light of this observation we tentatively refer to model (2) as mixture of Gaussian Bayesian network (MGBN).

The additional parameters β and σ^2 increase the complexity of a MGBN network compared to its underlying discrete BN. The i -th node with a_i categories contributes additional $1 + a_i$ parameters (one for σ_i^2 and a_i for $\beta'_{i,c}$), thus, increasing the overall complexity of $\mathcal{G}$ to

$$\alpha(\mathcal{G}) = \sum_{i=1}^n [1 + a_i + (a_i - 1) \prod_{j \in \Pi_i} a_j].$$

Example 2

Here we consider the expression levels of AKT 1 and GSK3A genes from the lung cancer data described and analyzed in Section 5 below. In Fig. 1, we show the cross plot for the original sample of size 62 (left), as well as for a sample of size 200, simulated from a fitted MGBN model (right). Also shown are the β parameters of the model corresponding to AKT 1 and GSK3A variables, drawn as vertical and horizontal lines, respectively. The cross points of the β -lines are centroids of Gaussian clusters. This is a situation where modeling continuous variables with categorical distributions, as the MGBN model does, seems appropriate.

4. EM Estimation

4.1. Problem Formulation

The main problem considered in this article is MGBN model estimation from observations on y . It involves learning the discrete BN $\mathcal{G}$ as well as estimating the mean parameters β and variances σ^2 . A standard solution is provided by the Maximum Likelihood principle. Moreover, the Expectation Maximization algorithm seems particularly well suited for the task because it can handle latent variables - in our case the discrete random vector x .

We assume that the network $\mathcal{G}$ has a known topological order and this is the index order, i.e. $x_1 < x_2 < \dots < x_n$. In many practical problems, such as those of recovering gene regulatory networks, this assumption is a natural one if viewed from the perspective that any prior

information for the node order should come from experimental studies. For example, the gene-node order in microarray studies can be inferred using perturbed samples. Identifying the causality relations between random variables only by statistical means is rarely possible for there might be many networks with different orders that fit the data equally well. Specifying an order resolves this issue and, eventually, makes the MLE network identifiable. This is the reason why in the considered here framework we provide MLE solutions only relative to given orders.

The EM algorithm is a standard technique for finding a local ML parameter configuration (Dempster, Laird & Rubin (1977)). It essentially requires optimization of the following score function of θ , $r(\theta|y, \theta^*) = El(\theta|y, x)$, where the expectation is taken with respect to the conditional distribution of $(x|y, \theta^*)$ for a fixed choice θ^* of the parameters. Starting from an initial θ^1 , a sequence θ^k is constructed according to $\theta^{k+1} = \arg \max_{\theta} r(\theta|\theta^k, y)$, until some convergence criteria are met.

Applying Eq. (4) we obtain $r(\theta|y, \theta^*) = \sum_{i=1}^n r_i(\theta_i|y, \theta^*)$, with r_i defined as

$$r_i(\theta_i|y, \theta^*) \equiv E_{x|y, \theta^*} \log P(y_i, x_i | x_{j \in \Pi_i}, \theta) = E_{x_i, x_{j \in \Pi_i} | y, \theta^*} \log P(y_i, x_i | x_{j \in \Pi_i}, \theta_i), \quad (6)$$

where $\theta = (\theta_i)_{i=1}^n$ and $\theta_i = (\beta_{i,k}, \sigma_i^2, \Pi_i, P(x_i | \Pi_i))$ is the model parameter corresponding to the i -th node. The last equality in (6) follows from the fact that $P(y_i, x_i | x_{j \in \Pi_i})$ depends on θ only through θ_i .

At the second step of EM, one needs to maximize $r(\theta|y, \theta^*)$ as a function of θ . Unfortunately, the expectations in (6) involve the calculation of the joint distribution of y 's. The latter problem can easily become computationally prohibitive since it requires

considering all possible configurations of x 's ($\prod_{i=1}^n a_i$ in total). Thus we exclude the possibility of a straightforward implementation of EM based on Eq. (6) as impractical.

In regard to the difficulties of finding $r(\theta|y, \theta^*)$, we observe that accessible seem to be only 'local' expectation calculations which involve distributions of limited subsets of nodes. For example, we can readily express the distribution of x_i and its parents conditional on y_i and $y_{j \in \Pi_i}$. Indeed, let us define $p_{i,c} \equiv P(x_i = c, \Pi_i = C | \mathcal{G})$, for $c \in C_i$ and $C \in C[\Pi_i]$, to denote the joint distribution of x_i and its parents Π_i in the network $\mathcal{G}$. Since y_j does not depend on $\mathcal{G}$ if the value of x_j is known, and all x_j depend on θ only through the network $\mathcal{G}$, we infer that

$$\begin{aligned} P(x_i = c, \Pi_i = C | y_i, y_{j \in \Pi_i}, \theta) &= \frac{P(x_i = c, \Pi_i = C | \mathcal{G}) P(y_i | x_i = c, \theta) \prod_{j \in \Pi_i} P(y_j | x_j = c_j, \theta)}{P(y_i, y_{j \in \Pi_i} | \theta)} \\ &= \frac{p_{i,c} \exp(-\frac{1}{2\sigma_i^2}(y_i - \beta_{i,c})^2 - \sum_{j \in \Pi_i} \frac{1}{2\sigma_j^2}(y_j - \beta_{j,c_j})^2)}{\sum_{d \in C_i} \sum_{D \in C[\Pi_i]} p_{i,d} \exp(-\frac{1}{2\sigma_i^2}(y_i - \beta_{i,d})^2 - \sum_{j \in \Pi_i} \frac{1}{2\sigma_j^2}(y_j - \beta_{j,d_j})^2)}, \end{aligned} \quad (7)$$

where $C = (c_j)_{j \in \Pi_i}$ and $D = (d_j)_{j \in \Pi_i}$. In addition, the conditional distribution of x_i is

$$P(x_i = c | y_i, y_{j \in \Pi_i}, \theta) = \sum_{C \in C[\Pi_i]} P(x_i = c, \Pi_i = C | y_i, y_{j \in \Pi_i}, \theta). \quad (8)$$

Because of the additive form of the score function, the maximization of $r(\theta|y, \theta^*)$ can be split into a sequence of n maximization problems, one for each θ_i . Motivated by this observation we propose to relax the baseline EM algorithm by considering the following

alternative score function $h(\theta|y, \theta^*) = \sum_{i=1}^n h_i(\theta_i|y, \theta^*)$, defined as

$$h_i(\theta_i|y, \theta^*) \equiv E_{x_i, x_{j \in \Pi_i} | y_i, y_{j \in \Pi_i}, \theta^*} \log P(y_i, x_i | x_{j \in \Pi_i}, \theta_i) = E_{x_i, x_{j \in \Pi_i} | y_i, y_{j \in \Pi_i}, \theta^*} [\log P(y_i | x_i, \beta_i, \sigma_i^2) + \log P(x_i | x_{j \in \Pi_i}, \theta_i)], \quad (9)$$

instead of the original (6). For each node, we thus consider the conditional expectation involving only those x 's and y 's used in the corresponding probability term in (6). Next, we briefly comment on the validity of using h instead of r as an EM score function.

Let $Y = \{y^1, \dots, y^m\}$ be a sample of y coming from a MGBN model with parameter θ_0 . Denote by $\bar{l}$, $\bar{r}$ and $\bar{h}$ the sample averages of the likelihood l , r and h ; for example,

$\bar{h}_i(\theta_i|Y, \theta^*) = \frac{1}{m} \sum_{j=1}^m h_i(\theta_i|y^j, \theta^*)$. Let $\hat{\theta}_{r,m}$ be a limit point of an EM sequence θ^k , $\theta^{k+1} = \arg \max_{\theta} r(\theta|\theta^k, y)$, and $\hat{\theta}_{h,m}$ be a limit point of an EM sequence based on h . Then, it must be that $\hat{\theta}_{r,m} = \arg \max_{\theta} \bar{r}(\theta|Y, \hat{\theta}_{r,m})$ and $\hat{\theta}_{h,m} = \arg \max_{\theta} \bar{h}(\theta|Y, \hat{\theta}_{h,m})$. We claim that the estimations based on r and h are asymptotically equivalent in the following sense: if $\hat{\theta}_{r,m}$ are consistent estimators of the true parameter θ_0 , so are $\hat{\theta}_{h,m}$. A support for this claim is given in Appendix A.

4.2. Algorithm Derivation

To evaluate the score function h , we start with the first term in (9)

$$E_{x_i | y_i, y_{j \in \Pi_i}, \theta^*} \log(P(y_i | x_i, \beta_i, \sigma_i^2)) = -\frac{1}{2} [\log(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2} E_{x_i | y_i, y_{j \in \Pi_i}, \theta^*} (y_i - \beta_{i,x_i})^2] = -\frac{1}{2} [\log(2\pi\sigma_i^2) + \frac{1}{\sigma_i^2} \sum_{c \in C_i} (y_i - \beta_{i,c})^2 q_{i,c}^*], \quad (10)$$

where $q_{i,c}^* = P(x_i = c | y_i, y_{j \in \Pi_i}, \theta^*)$ are the conditional probabilities defined by Eq. (8). For the second term in (9) we have

$$E_{x_i, x_{j \in \Pi_i} | y_i, y_{j \in \Pi_i}, \theta^*} \log P(x_i | x_{j \in \Pi_i}) = \sum_{C \in C[\Pi_i]} \sum_{c \in C_i} \log P(x_i = c | \Pi_i = C) q_{i,c}^*, \quad (11)$$

where $q_{i,c}^* \equiv P(x_i = c, \Pi_i = C | y_i, y_{j \in \Pi_i}, \theta^*)$ as in Eq. (7).

At the second step of the EM algorithm, we need to maximize the sum of (10) and (11) in order to find an optimal network $\mathcal{G}$. This problem requires for each i to find, first, the optimal parenthood Π_i of node x_i and second, the conditional probability $P(x_i | \Pi_i)$. Provided Π_i is known, the optimal conditional probability can be easily derived. Indeed, by using the fact that for every fixed $C \in C[\Pi_i]$

$$\arg \max_v \left\{ \sum_c \log(v_c) q_{i,c}^* \mid \sum_c v_c = 1 \right\} = \left(\frac{q_{i,c}^*}{\sum_d q_{i,d}^*} \right),$$

for the expectation in Eq. (11) we can write

$$\max_P E_{x_i, x_{j \in \Pi_i} | y_i, y_{j \in \Pi_i}, \theta^*} \log P(x_i | x_{j \in \Pi_i}) = \sum_C \sum_c q_{i,c}^* \log(q_{i,c}^* / \sum_d q_{i,d}^*). \quad (12)$$

Recall that the EM algorithm is an iterative procedure that for a sample Y and a starting triple $\theta^0 = (\sigma^0, \beta^0, \varphi^0)$, constructs a sequence $\theta^k = (\sigma^k, \beta^k, \varphi^k)$, $k = 1, 2, \dots$, until a target convergence threshold for $\tilde{h}(\theta^{k+1} | Y, \theta^k)$ is achieved. To simplify the algorithm iteration equations, we introduce notations for the probabilities (7) and (8) referring to iteration index k and sample instance l

$$q_{i,c}^{k,l} \equiv \frac{p_{i,c}^k \exp(-\frac{1}{2(\sigma^k)^2}[(y_i^l - \beta_{i,c}^k)^2 + \sum_{j \in \Pi_i} (y_j^l - \beta_{j,c}^k)^2])}{\sum_{d \in C_i} \sum_{D \in C[\Pi_i]} p_{i,d}^k \exp(-\frac{1}{2(\sigma^k)^2}[(y_i^l - \beta_{i,d}^k)^2 + \sum_{j \in \Pi_i} (y_j^l - \beta_{j,d}^k)^2])},$$

$$q_{i,c}^{k,l} \equiv \sum_{D \in C[\Pi_i]} q_{i,cD}^{k,l}, \text{ and let } \bar{q}_{i,c}^k \equiv \sum_{l=1}^m q_{i,c}^{k,l},$$

where $p_{i,c}^k$ and $p_{i,cD}^k$ denote the probabilities $P(x_i | G^k)$ and $P(x_i, \Pi_i | G^k)$, respectively. We consider $q_{i,c}^{k,l}$ and $q_{i,cD}^{k,l}$ as functions of the parent set Π_i .

Then, at the k -th iteration, the EM algorithm computes θ^{k+1} from θ^k as a solution of

$$\theta^{k+1} \equiv (\sigma^{k+1}, \beta^{k+1}, \mathcal{G}^{k+1}) = \arg \max_{\sigma, \beta, \mathcal{G}} \sum_{i=1}^n H_i^k(Y, \theta_i), \quad (13)$$

where, by utilizing Eq. (12) for the sums $\bar{q}_{i,c}^k$, we define $H_i^k(Y, \theta_i)$ to be

$$-\frac{m}{2} \log(2\theta\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{l=1}^m \sum_{c \in C_i} q_{i,c}^{k,l} (y_i^l - \beta_{i,c}^k)^2 + \sum_{C \in C[\Pi_i]} \sum_{c \in C_i} \bar{q}_{i,c}^k \log(\bar{q}_{i,c}^k / \sum_{d \in C_i} \bar{q}_{i,d}^k).$$

Note that the optimization function in (13) is additive for the nodes of the network and $\theta_i^{k+1} = \arg \max_{\theta_i} H_i^k(Y, \theta_i)$. This fact greatly simplifies the computation process for allowing the evaluation of σ^{k+1} , β^{k+1} and G^{k+1} on node by node basis.

Without imposing some restrictions on network $\mathcal{G}^{k+1}$, Eq. (13) will always select the maximum connected network - one for which each node has as parents all the nodes appearing earlier in the node order. This trivial solution is the least economical one and surely will over-fit any reasonable data. Apparently, some model selection restrictions on $\mathcal{G}^{k+1}$ are necessary such as searching for networks with fixed in advance complexity. Alternatively, we may apply some standard model selection criteria such as AIC or BIC. Then the complexity of $\mathcal{G}^k$ may vary from iteration to iteration. Next, we write the estimation equations for these cases.

4.3. EM Algorithm with Model Selection

If the parent set sizes $|\Pi_i|$ of the nodes are known and fixed, then it is straightforward to write the iteration equations for θ^{k+1} . In this case the complexities of the resulting networks are predetermined. The optimal graph structure $\mathcal{G}$ in (13) can be found following the topological order of the nodes - the parenthood of the i -th node is a subset of size $|\Pi_i|$ of nodes appearing earlier in the order, that maximizes H_i^k and can be searched for exhaustively. The optimal β^* 's and σ^* 's are the usual MLE solutions from the linear regression analysis. Specifically, the solution of (13) includes the following update rules for β_i^k, σ_i^k and Π_i^k

$$\beta_{i,c}^{k+1} = \sum_{l=1}^m q_{i,c}^{k,l} y_i^l / \bar{q}_{i,c}^k, \quad (14)$$

$$(\sigma_i^{k+1})^2 = \frac{1}{m} \sum_{l=1}^m \sum_{c \in C_i} q_{i,c}^{k,l} (y_i^l - \beta_{i,c}^{k+1})^2, \quad (15)$$

$$\Pi_i^{k+1} = \underset{\Pi_i}{\operatorname{argmax}} \left\{ -\frac{m}{2} \log \left(\sum_{l=1}^m \sum_{c \in C_i} q_{i,c}^{k,l} (y_i^l - \beta_{i,c}^{k+1})^2 \right) + \sum_{C \in C[\Pi_i]} \sum_{c \in C_i} \bar{q}_{i,c}^k \log(\bar{q}_{i,c}^k / \sum_{d \in C_i} \bar{q}_{i,d}^k) \right\} \quad (16)$$

where Π_i is chosen among the subsets of $\{1, \dots, i-1\}$ of size $|\Pi_i|$. At every iteration step k , the update equations (14)–(16) are applied in order from x_1 to x_n , or according to the node topological order if it is different from the assumed one $x_1 < x_2 < \dots < x_n$.

If the parent set sizes are unknown, one needs an additional selection criterion to avoid the trivial, most connected, network as a solution of (13). This is essentially a model selection problem similar to that considered in Salzmann & Almudevar (2006). A natural choice is a network with a specified, target, complexity A . Then the algorithm guarantees optimally for this target complexity. To express it formally we modify Eq. (13) to

$$(\sigma^{k+1}, \beta^{k+1}, \mathcal{G}^{k+1}) = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^n H_i^k(Y, \theta_i) \mid \alpha(\mathcal{G}) = A \right\}. \quad (17)$$

A second option is EM algorithm with AIC network selection, which replace (13) with

$$(\sigma^{k+1}, \beta^{k+1}, \mathcal{G}^{k+1}) = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^n H_i^k(Y, \theta_i) - \alpha(\mathcal{G}) \right\}, \quad (18)$$

where, recall, $\alpha(\mathcal{G})$ is the network complexity. A third option is to use the BIC criterion that leads to the following iteration equation

$$(\sigma^{k+1}, \beta^{k+1}, \mathcal{G}^{k+1}) = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^n H_i^k(Y, \theta_i) - 0.5\alpha(\mathcal{G})\log(m) \right\}. \quad (19)$$

If n is the number of nodes and $\max c_i = \max_{i=1}^n c_i$ is the maximum number of categories per node, the complexity of one EM iteration is exponential to the maximum parent size $\max p = \max_{i=1}^n |\Pi_i|$ and is of order $O(m * (n * \max c)^{\max p + 1})$. Note that the computational complexity of the algorithm, both in processing time and memory, may quickly become prohibitive with the increase of $\max p$. Nevertheless, estimating networks with $n \leq 100$, $\max c \leq 3$ and $\max p \leq 3$ is currently quite manageable (see the simulation results in the next section) and these are sufficient for many practical problems.

5. Results

5.1. Simulation Studies

Here we demonstrate the performance of the proposed EM learning algorithm on simulated networks. In each test trial we generate a random MGBN, serving as a ground truth network, then draw samples of varying size from it and finally, do estimation. The network structure recovery is measured by the so called F -score, which is the harmonic mean of the specificity ($TN/(TN + FP)$) and sensitivity ($TP/(TP + FN)$), where TP (true positives) is the number of correctly found directed edges, FP is the number of edges present in the estimated network but not in the original one, and FN is the number of edges present in the original network but not in the estimated one. A F -score of 1 means perfect reconstruction. MGBN is compared to another network learning with the discretization and network fitting steps performed separately. The latter method is denoted as D+BN. Note that without some assumptions on the marginal distributions of the nodes, the probability structure of a discrete BN can not be recovered, although the edge structure eventually can. The MGBN model, being a parametric one, provides such set of assumptions. Since D+BN with a general discretization method can not be properly compared to MGBN, to place both methods on equal footing, we perform the discretization step in D+BN using the (known) node marginal distributions from the original networks.

Table 2 presents results for two networks with sizes 25 and 100, 3 categories per node and a maximum of three parents per node. The networks have β -parameters in the $[-1, 1]$ range and a substantial noise level with $\sigma^2 = 0.1$. From each of the networks four samples of size 100, 200, 400 and 800 are drawn and MGBN and D+BN estimation performed. For each trial we report the F -scores as well as the processing times in seconds. We limit the maximum parent size of the estimated MGBN to 2, one less than that of the original, which prevent perfect reconstruction but speeds up the computations. The number of EM-steps for MGBN varies from trial to trial, but is limited to a maximum of 60.

Based on the F -score numbers, MGBN performs better than D+BN, although with much higher computational cost. The results also show that the computational time of MGBN estimation is according to the theoretical $O(m(3n)^3)$ -complexity and is linear on the sample size.¹ Note that the presented simulation results are in favor of the D+BN method, for in a real problem, the node marginal distributions will be unknown and any other discretization method will result in worse performance of D+BN.

¹The simulations are run on 2.66 GHz, i5 processor. D+BN and MGBN estimations are performed using *catnet* and *mugnet R* packages.

5.2. Data Example: Learning Gene Regulatory Network

In this section we illustrate the application potential of the MGBN model by a small microarray data study. The renewed interest in the graphical statistical models in recent years is due to the expectation that they would help understanding the complicated mechanism of gene and protein interactions within cells. The data we have chosen is from one of the BROAD Institute GSEA's publicly available datasets² of transcriptional profiles from lung cancer cell lines, the Boston cohort. Extensive gene expression analysis of the data is published in Beer et al. (2002) and Bhattacharjee et al. (2001). The lung dataset contains about 12600 genes and 62 samples taken from patients with similar disease condition. Furthermore, the sample is separated into two equal classes of size 31, depending on the survival status at the end of the study - 'alive' and 'deceased'.

For illustration purposes only, we concentrate attention on a small subset of genes and try to learn the interaction between them using the framework we have just presented. We select the following nine genes: *IGF1*, *PIK3CA*, *PT EN*, *PDK1*, *AKT1*, *BAD*, *CASP9*, *GSK3A* and *MT OR*, that play an important role in the *PI3K/PT EN/AKT* pathway. This pathway is believed to have a critical role in cell survival process and its activation is linked to tumor genesis, Saal et al. (2007). Briefly, *IGF1* is a growth factor gene, *PIK3CA* encodes the protein *p110α* and is mutated in a range of human cancers, *PT EN* is another oncogene with role in tumor suppression, and *AKT1* plays anti-apoptotic action when activated. The selected subset of genes is ordered following the available in the literature gene interaction causality.

We fit MGBN to the combined data of the two survival classes and then look at the class differences using the fitted model. To avoid over-fitting the data, we carefully choose the number of node categories and the maximum number of parents per node such that the maximum node complexity is smaller than the sample size by some factor. With 2 parents and 2 categories per node the maximum node complexity is 7. In this case, the saturated network with 1 parent per node has complexity 44, while that with 2 parents has complexity 58. We fit MGBN using three different network selection criteria: AIC, 1-parent saturated network and 2-parents saturated network. The found three networks, *G1*, *G2* and *G3*, are shown in Fig. 2.

A distinguishing feature of any graphical probability model is its accessible interpretation. For example, the AIC-selected network, *G1*, includes the top five most significant arcs, with *PIK3CA* → *PT EN* having the highest likelihood score. *G1* is a sub-network of *G2*, which is saturated with all gene-nodes, except the root, having one parent - their most significant predictors. In the third network, *G3*, each parent pair is the most significant among the possible ones for the corresponding node. The found three networks confirm some known gene regulations such as *IGF1* → *PIK3CA* → *PT EN* → *PDK1* and *AKT1* → *GSK3A*.

The conditional probabilities of the fitted networks can be used to reveal details about the difference between the two survival classes. In network *G3*, *BAD* depends on *PT EN* and *AKT1*. To quantify the class difference, we estimate the *BAD* conditional distribution using the 'alive' and 'deceased' sub-samples. The obtained two empirical conditional distributions are plotted in Fig. 3. Recall that each discrete gene-node has two categories (1 and 2) and thus, the pair (*PT EN*, *AKT1*) has four possible configurations. When *AKT1* is down-regulated (*AKT1* = 1), the distributions of *BAD* are essentially the same for the two classes and are not influenced by *PT EN*. However, when *AKT1* is up-regulated (*AKT1* = 2), the regulation of *BAD* depends on *PT EN* and is different between the survival classes. If both *PT EN* and *AKT1* are up-regulated (see the fourth column in Fig. 3), the activity of *BAD* is

²www.broadinstitute.org/gsea/datasets.jsp

slightly disturbed in the ‘deceased’ class in comparison to the ‘alive’ one. Similar plots, Fig. 4, show the behavior of *AKT1*. A consistent up-regulation of *AKT1* is evident for class ‘deceased’, in contrast to class ‘alive’, where its activity is moderated by *PIK3CA* - see the second column, *IGF1* = 1, *PIK3CA* = 2. These observations are consistent with the inhibiting role of *AKT1* in cell apoptosis, which in turn, as it is known, leads to cancer cell proliferation.

The MGBN model is implemented in the *mugnet* R-package and is available from CRAN repositories. The lung cancer data and analysis in this section are also available as demonstration in *mugnet*.

6. Conclusion

We have introduced the MGBN model - a discrete Bayesian network with Gaussian errors distributed independently across its nodes. It provides a framework for representing continuous variables via unconstrained categorical probability distributions. The model allows unification of two so far independently considered problems - data discretization and network learning. We have also derived an efficient procedure for MGBN estimation based on the Expectation Maximization method. In the appendix we briefly discuss the convergence properties of our estimator, including the MLE consistency, and present directions for more detailed treatment of this important subject. We believe that the MGBN model is suited for a variety of problems that involve complicated dependencies, such as learning gene/protein interactions from microarray data, as our small study has illustrated. The framework can be further extended by replacing the Gaussian error model with other distributions from the exponential family, thus diversifying its application scope.

Acknowledgments

This work was supported by NIH grant K99LM009477 from the National Library Of Medicine. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Library Of Medicine or the National Institutes of Health.

The author thanks Peter Salzman for many valuable discussions and suggestions. The author also thanks the reviewers, whose careful reading and valuable comments have helped for the improvement of this paper.

References

- Bhattacharjee A, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;98(24):13790–13795. [PubMed: 11707567]
- Beer D, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma, 2002. *Nat Med* 2002 Aug;8(8):816–24. Jul 15. [PubMed: 12118244]
- Buntine WL. A Guide to the Literature on Learning Probabilistic Networks from Data. *IEEE Transactions on Knowledge and Data Engineering* 1996;8(2)
- Chickering, DM. Learning from data: Artificial intelligence and statistics V. Springer-Verlag; 1996. Learning Bayesian networks is NP-complete; p. 121130
- Chickering DM, Heckerman D. Efficient Approximations for the Marginal Likelihood of Bayesian Networks with Hidden Variables. *Machine Learning* 1996;29:181212.
- Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B* 1977;39:1–38.
- Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*. 1995
- Heckerman D, Geiger D. Learning Bayesian Networks: A Unification for Discrete and Gaussian Domains. *Proceedings of UAI '95* 1995:274–284.

- Lauritzen SL. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* 1995;19:191–201.
- Lauritzen, SL. *Journal of the American Statistical Association*. 1992. Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models.
- Lauritzen, SL.; Wermuth, N. Tech Rep R-84-8. Institute of Electronic Systems, Aalborg University; Denmark: 1984. Mixed interaction models.
- Lauritzen SL, Jensen F. Stable local computation with conditional Gaussian distributions. *Statistics and Computing* 2001;191–203.
- Friedman, N. Learning Belief Networks in the Presence of Missing Values and Hidden Variables. *Proceedings of the Fourteenth International Conference on Machine Learning*; 1997. p. 125-133.
- Neapolitan, RE. *Learning Bayesian Networks*. Prentice Hall; 2003.
- Saal L, et al. Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity, 2007. *Proc Natl Acad Sci U S A* 2007 May 1;104(18): 7564–9. Apr 23. [PubMed: 17452630]
- Salzman P, Almudevar A. Using Complexity for the Estimation of Bayesian Networks. *Statistical Applications in Genetics and Molecular Biology* 2006;5(1)

Appendix A

As in the main text, let $Y = \{y^1, \dots, y^m\}$ be a sample from a MGBN with true parameter θ_0 , $\bar{r}$ and $\bar{h}$ be the sample averages of r and h , and $\hat{\theta}_{r,m}$ and $\hat{\theta}_{h,m}$ be limit points of r and h -based EM sequences, respectively. By introducing the complement index set $\bar{i}, \bar{\Pi}_i = \{j: j \neq i \text{ and } j \notin \Pi_i\}$ and observing that

$$h_i(\theta_i|y, \theta^*) = E_{y_{\bar{i}, \bar{\Pi}_i}}[y_i, y_{j \in \Pi_i} r_i(\theta_i|y, \theta^*)], \quad (\text{A.1})$$

we immediately obtain $E_{y|\theta^*} r_i(\theta_i|y, \theta^*) = E_{y|\theta^*} h_i(\theta_i|y, \theta^*)$, for all θ and θ^* . Therefore, by the strong law of large numbers, both $\bar{r}(\theta|Y, \theta^*)$ and $\bar{h}(\theta|Y, \theta^*)$ converge to $E_{r(\theta|y, \theta^*)}$, a.s., provided the expectations with respect to $y|\theta^*$ exist. Tentatively, we expect that the estimators $\hat{\theta}_{r,m}$ and $\hat{\theta}_{h,m}$ share the same properties.

The question of consistency of MLE for the MGBN model, which does not unconditionally hold, is more involved to be discussed in details here and we will only sketch one approach to this problem. For simplicity, hereafter we will assume that the parent sizes, $|\Pi_i|$, of all nodes are fixed. Any other setting would involve more complex structure for the parameter space and, generally, would require more stringent conditions to guarantee the uniqueness of MLE.

In order to show the asymptotic equivalence of the h and r -based EM estimations, we will assume the following consistency sufficient conditions on r known from the general asymptotic theory. Let there be a parameter set Θ , with θ_0 in the interior of Θ , such that

$$P(\hat{\theta}_{r,m} \in \Theta) \rightarrow 1 \text{ and } P(\hat{\theta}_{h,m} \in \Theta) \rightarrow 1, \text{ as } m \rightarrow \infty, \quad (\text{A.2})$$

$$\sup_{\theta^* \in \Theta} E_{y|\theta^*} \sup_{\theta \in \Theta} |\bar{r}_i(\theta_i|Y, \theta^*) - E_{y|\theta^*} r_i(\theta_i|y, \theta^*)| \rightarrow 0, \text{ as } m \rightarrow \infty, \quad (\text{A.3})$$

for $i = 1, \dots, n$, and that all $Er_i(\theta_i|y, \theta_0)$ have well separated maximums at $\theta_{0,i}$. The latter guarantees that the system of equations $\theta_i^* = \arg\max_{\theta_i} Er_i(\theta_i|y, \theta^*)$ has a unique solution for $\theta^* = \theta_0$. If we let $\theta^* = \theta_0$ in condition (A.3), we obtain uniform convergence in mean of the sample log-likelihood of y to the expected one, which, together with the third condition, the well separateness of the maximums, is sufficient for the MLE consistency.

Next we observe that

$$\sup_{\theta \in \Theta} |\bar{h}_i(\theta_i|Y, \theta^*) - Eh_i(\theta_i|y, \theta^*)| \leq \sup_{\theta \in \Theta} E_{y_{i, \Pi_i} | y_i, y_{j \in \Pi_i}} |\bar{r}_i(\theta_i|Y, \theta^*) - Er_i(\theta_i|y, \theta^*)| \leq E_{y_{i, \Pi_i} | y_i, y_{j \in \Pi_i}} \sup_{\theta \in \Theta} |\bar{r}_i(\theta_i|Y, \theta^*) - Er_i(\theta_i|y, \theta^*)|$$

by (A.1). Then condition (A.3) holds for h as well and therefore, the score function h also produces consistent estimators.

Are the above consistency sufficient conditions reasonable to be assumed? Which ones are general model properties that can be induced and which ones are essential? The identifiability of MLE, guaranteed by the well separateness of the maximum of $Er(\theta|y, \theta_0)$, is an essential condition and it is model specific - it can not be induced. Let Θ be a bounded compact subset of the parametric space that contains θ_0 in its interior and that does not intersect any hyper-plane $\sigma_i = 0$ and $P(x_i = c | \Pi_i = C, \theta_i) = 0$. Such Θ exists whenever the conditional distributions in θ_0 are away from zero. Then, it can be shown that condition (A.2) follows from the well-separateness of the maximum of $Er_i(\theta_i|y, \theta_0)$ and the compactness of Θ . To show (A.3) we express (6) as

$$r_i(\theta_i|y, \theta^*) = -\frac{1}{2} \log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{c \in C_i} (y_i - \beta_{i,c})^2 P(x_i = c | y, \theta^*) + \sum_{c \in C_i, C \in C[\Pi_i]} \log(P(x_i = c | \Pi_i = C, \theta_i)) P(x_i = c, \Pi_i = C | y, \theta^*)$$

and observe that, because of the compactness of Θ , the supremum in (A.3) is bounded by

$$\max_{\theta^*} E_{y|\theta^*} \max_{\theta} \{ \overline{b_1 | (y_i - \beta_i)^2} - E(y_i - \beta_i)^2 + b_2 | \overline{P(x_i, \Pi_i | y, \theta^*)} - EP(x_i, \Pi_i | y, \theta^*) \},$$

for $b_1 = \max_{\Theta} (1/2\sigma_i^2) < \infty$ and $b_2 = \max_{\Theta} \log(P(x_i = c | \Pi_i = C, \theta_i)) < \infty$, where the bar denotes sample average. Further considerations show that both terms above converge to zero. The convergence of the first one can be verified using the properties of the Gaussian distribution. In the second, since there are only finitely many possible parent sets Π_i , in fact at most

$$\binom{n-1}{|\Pi_i|}, \text{ we have}$$

$$\max_{\theta} |\overline{P(x_i, \Pi_i | y, \theta^*)} - EP(x_i, \Pi_i | y, \theta^*)| \leq \sum_{\Pi_i} |\overline{P(x_i, \Pi_i | y, \theta^*)} - EP(x_i, \Pi_i | y, \theta^*)|,$$

and the latter sum converges to zero in mean by an application of the Dominated convergence theorem. Therefore, for Θ as chosen above, (A.3) is satisfied.

Figure 1. The cross plot of *GSK3A* and *AKT1* genes for an original (left) and simulated (right) data. The latter is obtained from a MGBN model fitted to the original sample. See Example 2 and the analysis in Section 5 for more details.

Figure 2. Three networks representing the lung cancer data according to different selection criteria. See the main text for details.

Figure 3.

The probabilities of *BAD* conditional on *PTEN* and *AKT1* for class 'alive', first row, and 'deceased', second row. The columns in each row correspond to the four combinations of *PTEN* and *AKT1*. The largest difference is observed in the second column, corresponding to $PTEN = 1$ and $AKT1 = 2$. There is also a slight difference when $PTEN = 2$ and $AKT1 = 2$ (fourth column).

Figure 4.

The probabilities of *AKT* conditional on *IGF1* and *PIK3CA* for class 'alive', first row, and 'deceased', second row. For the latter we observe a consistent activity of *AKT1*. There is an apparent class difference when *IGF1* = 1 and *PIK3CA* = 2 (second column).

Table 1

A simple 2-node MGBN model, $A \rightarrow B$, with its discrete probability table (on the left) and a sample drawn from it (on the right).

Table 2

Performance comparison of MGBN vs. separate discretization and BN fitting (D+BN), both using AIC selection. Reported are the F -scores and the processing time in seconds (in brackets).

m	100	200	400	800
Network with $n = 25$ nodes and 24 edges				
D+BN	0.28 (1.1)	0.54 (1.1)	0.73 (1.1)	0.87 (1.2)
MGBN	0.34 (35)	0.73 (62)	0.83 (98)	0.88 (178)
Network with $n = 100$ nodes and 89 edges				
D+BN	0.23(54)	0.48 (57)	0.63 (58)	0.77 (62)
MGBN	0.37(1635)	0.62 (2990)	0.74 (5523)	0.85 (11048)