Learning Bayesian Networks with Mixed Variables

Susanne Gammelgaard Bøttcher
Department of Mathematical Sciences
Aalborg University
9220 Aalborg, Denmark

Abstract

The paper considers conditional Gaussian networks. As conjugate local priors, we use the Dirichlet distribution for discrete variables and the Gaussian-inverse Gamma distribution for continuous variables, given a configuration of the discrete parents. We assume parameter independence and complete data. Further, the network-score is calculated. We then develop a local master prior procedure, for deriving parameter priors in CG networks. The local master procedure satisfies parameter independence, parameter modularity and likelihood equivalence.

1 Introduction

The aim of this paper is to present a method for learning the parameters and the structure of a Bayesian network with discrete and continuous variables. In Heckerman, Geiger & Chickering (1995) and Geiger & Heckerman (1994) this was done for respectively discrete networks and Gaussian networks.

We define the local probability distributions such that the joint distribution of the random variables is a conditional Gaussian (CG) distribution. We do not allow discrete variables to have continuous parents, so the network factorizes into a discrete part and a mixed part. The local conjugate parameter priors are for the discrete part of the network specified as Dirichlet distributions and for the mixed part of the network as Gaussian-inverse Gamma distributions for each configuration of discrete parents.

To learn the structure of a CG network, we find the network-score $p(d, D)$. Further, we derive a method for finding the prior distribution of the parameters in possible structures, from marginal priors calculated from an imaginary database. The method satisfies parameter independence, parameter modularity and likelihood equivalence. Further, if used on networks with only discrete or continuous variables, it coincides with the methods developed in Heckerman et al. (1995) and Geiger & Heckerman (1994).

2 Bayesian networks

A Bayesian network is a graphical model that encodes the joint probability distribution for a set of variables. For terminology and theoretical aspects on graphical models, see Lauritzen (1996). In this paper we define it as

- A Directed Acyclic Graph (DAG) $D = (V, E)$, where V is a finite set of vertices and E is a finite set of directed edges between the vertices. The DAG defines the structure of the Bayesian network.
- To each vertex $v \in V$ in the graph corresponds a random variable X_v. The set of variables associated with the graph D is then $X = (X_v)_{v \in V}$. Often we do not distinguish between a variable X_v and the corresponding vertex v.
- To each vertex v with parents $\mathrm{pa}(v)$, there is attached a local probability distribution, $p(x_v|x_{\mathrm{pa}(v)})$. The set of local probability distributions for all variables in the network is denoted $\mathcal{P}$.
- The possible lack of directed edges in D encodes these conditional independencies between the random variables X through the factorization of the joint probability distribution,

$$p(x) = \prod_{v \in V} p(x_v|x_{\mathrm{pa}(v)}). \tag{1}$$

A Bayesian network for a set of random variables X is thus the pair $(D, \mathcal{P})$. In order to specify a Bayesian network, we must therefore specify a DAG D and a set $\mathcal{P}$ of local probability distributions.

3 Bayesian networks for mixed variables

In this paper we are interested in specifying networks for random variables X of which some are discrete and some are continuous (qualitative/quantitative). So we consider a DAG $D = (V, E)$ with variables/vertices $V = \Delta \cup \Gamma$, where Δ and Γ are the sets of discrete and continuous variables, respectively. The corresponding random variables X can then be denoted $X = (X_v)_{v \in V} = (I, Y) = ((I_\delta)_{\delta \in \Delta}, (Y_\gamma)_{\gamma \in \Gamma})$, i.e. we use I and Y for the sets of discrete and continuous variables respectively. We denote the set of levels for each discrete variable $\delta \in \Delta$ as $\mathcal{I}_\delta$.

In this paper we do not allow discrete variables to have continuous parents. This is to ensure availability of exact local computation methods, see Lauritzen (1992) and Lauritzen & Jensen (1999). So the set of edges E satisfies $E \subseteq (\Gamma \times \Delta)^{\complement}$, where $\complement$ denotes the complement. Now we need to specify the set of local probability distributions $\mathcal{P}$. As we have no discrete children of continuous parents, the joint probability distribution factorizes as follows:

$$p(x) = p(i, y) = \prod_{\delta \in \Delta} p(i_\delta | i_{\mathrm{pa}(\delta)}) \prod_{\gamma \in \Gamma} p(y_\gamma | i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}).$$

Note that $i_{\mathrm{pa}(\gamma)}$ and $y_{\mathrm{pa}(\gamma)}$ denote observations of the discrete and continuous parents respectively, i.e. $i_{\mathrm{pa}(\gamma)}$ is an abbreviation of $i_{\mathrm{pa}(\gamma) \cap \Delta}$ etc.

We see that the joint probability distribution factorizes into a purely discrete part and a mixed part. First we look at the discrete part.

3.1 The discrete part of the network

We assume that the local probability distributions are just unrestricted discrete distributions with

$$p(i_\delta | i_{\mathrm{pa}(\delta)}) \geq 0 \quad \forall \quad \delta \in \Delta.$$

A way to parameterize this is to say that

$$\theta_{i_\delta | i_{\mathrm{pa}(\delta)}} = p(i_\delta | i_{\mathrm{pa}(\delta)}, \theta_{\delta | i_{\mathrm{pa}(\delta)}}), \tag{2}$$

where $\theta_{\delta | i_{\mathrm{pa}(\delta)}} = (\theta_{i_\delta | i_{\mathrm{pa}(\delta)}})_{i_\delta \in \mathcal{I}_\delta}$.

Furthermore $\sum_{i_\delta \in \mathcal{I}_\delta} \theta_{i_\delta | i_{\mathrm{pa}(\delta)}} = 1$ and $0 \leq \theta_{i_\delta | i_{\mathrm{pa}(\delta)}} \leq 1$.

So using this parameterization, the discrete part of the joint probability distribution is given by

$$p(i | (\theta_{\delta | i_{\mathrm{pa}(\delta)}})_{\delta \in \Delta}) = \prod_{\delta \in \Delta} p(i_\delta | i_{\mathrm{pa}(\delta)}, \theta_{\delta | i_{\mathrm{pa}(\delta)}}). \tag{3}$$

3.2 The mixed part of the network

Now we consider the mixed part. We assume that the local probability distributions are Gaussian linear regressions on the continuous parents, with parameters depending on the configuration of the discrete parents. So let the parameters in the distribution be given by $\theta_{\gamma | i_{\mathrm{pa}(\gamma)}} = (f_{\gamma | i_{\mathrm{pa}(\gamma)}}, \beta_{\gamma | i_{\mathrm{pa}(\gamma)}}, \sigma^2_{\gamma | i_{\mathrm{pa}(\gamma)}})$. Then

$$(Y_\gamma | i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}, \theta_{\gamma | i_{\mathrm{pa}(\gamma)}}) \sim \mathcal{N}(f_{\gamma | i_{\mathrm{pa}(\gamma)}} + \beta_{\gamma | i_{\mathrm{pa}(\gamma)}} y_{\mathrm{pa}(\gamma)}\ , \sigma^2_{\gamma | i_{\mathrm{pa}(\gamma)}}), \tag{4}$$

where $\beta_{\gamma | i_{\mathrm{pa}(\gamma)}}$ are the regression coefficients, $f_{\gamma | i_{\mathrm{pa}(\gamma)}}$ is the conditional mean, and $\sigma^2_{\gamma | i_{\mathrm{pa}_\gamma}}$ is the conditional variance. The mixed part of the joint distribution can now be written as

$$p(y | i, (\theta_{\gamma | i_{\mathrm{pa}(\gamma)}})_{\gamma \in \Gamma}) = \prod_{\gamma \in \Gamma} p(y_\gamma | i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}, \theta_{\gamma | i_{\mathrm{pa}(\gamma)}}). \tag{5}$$

Further, the joint probability distribution $p(i, y | \theta)$, where

$$\theta = ((\theta_{\delta | i_{\mathrm{pa}(\delta)}})_{i_{\mathrm{pa}(\delta)} \in \mathcal{I}_{\mathrm{pa}(\delta)}}, (\theta_{\gamma | i_{\mathrm{pa}(\gamma)}})_{i_{\mathrm{pa}(\gamma)} \in \mathcal{I}_{\mathrm{pa}(\gamma)}})$$

is given by the product of (3) and (5). Notice that when the local probability distributions are given by (2) and (4), the joint probability distribution for X is a CG distribution (conditional Gaussian) with density of the form

$$p(i) |2\pi \Sigma_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y - m_i)^T \Sigma_i^{-1} (y - m_i)\}.$$

For each i, m_i is the unconditional mean, that is unconditional on continuous variables and Σ_i is the covariance matrix for the continuous variable in the network. In Shachter & Kenley (1989) formulas for calculating Σ_i from the local probability distributions can be found. A Bayesian network where the joint probability distribution is a CG distribution is in the following called a CG network.

4 Learning the parameters in a CG network

When constructing a Bayesian network, there is, as mentioned in Section 2, two things to consider, namely specifying the DAG and specifying the local probability distributions. In this section we assume that the structure of the DAG is known and the distribution type is determined as in the previous section. Now we consider the specification of the parameters in the distributions.

4.1 Some simplifying properties

Here we define the prior distributions of the parameters such that they are conjugate for the observations in question. Further, we assume that the parameters associated with one variable is independent of the parameters associated with the other variables. This assumption was introduced by Spiegelhalter & Lauritzen (1990) and we denote it global parameter independence. In addition to this, we will assume that the parameters are independent for each configuration of the discrete parents, which we denote as local parameter independence. So if the parameters have the property of global parameter independence and local parameter independence, then

$$
\begin{aligned}
p(\theta) = &\prod_{\delta \in \Delta} \prod_{i_{\mathrm{pa}(\delta)} \in \mathcal{I}_{\mathrm{pa}(\delta)}} p(\theta_{\delta|i_{\mathrm{pa}(\delta)}}) \\
&\times \prod_{\gamma \in \Gamma} \prod_{i_{\mathrm{pa}(\gamma)} \in \mathcal{I}_{\mathrm{pa}(\gamma)}} p(\theta_{\gamma|i_{\mathrm{pa}(\gamma)}}),
\end{aligned}
\tag{6}
$$

and we will refer to (6) simply as parameter independence.

A consequence of parameter independence is that, for each configuration of the discrete parents, we can update the parameters in the local distributions independently. This also means, that if we have local conjugacy, i.e. the distribution of $\theta_{\delta|i_{\mathrm{pa}(\delta)}}$ and $\theta_{\gamma|i_{\mathrm{pa}(\gamma)}}$ belongs to a conjugate family, then because of parameter independence, we have global conjugacy, i.e. the distribution of $p(\theta)$ belongs to a conjugate family. Further we will assume that the database d of cases, from which the parameters are updated, is complete, i.e. we have no missing observations. Due to parameter independence, the factorizations in (3) and (5), and the assumption of complete data, the parameters stay independent given data. We call this property posterior parameter independence. In other words, the properties of local and global independence are conjugate.

4.2 Learning in the discrete case

In the discrete part of the network we assumed that the local probability distributions are unrestricted discrete distributions defined as in (2). As pointed out in the previous section we can, because of the assumption of parameter independence, find the posterior distribution of $\theta_{\delta|i_{\mathrm{pa}(\delta)}}$ for each δ and each configuration of $\mathrm{pa}(\delta)$ independently.

Let $x^c \in d$ be a case in a database $d = \{x^1, \ldots, x^n\}$, where the configuration of the parents is $i^c_{\mathrm{pa}(\delta)}$. As the network can be partitioned in a pure discrete part and a mixed part, we can just consider the discrete part of the case, namely i^c.

A conjugate family for observations from (2), is the family of Dirichlet distributions. Let the prior distribution of $\theta_{\delta|i^c_{\mathrm{pa}(\delta)}}$ be a Dirichlet distribution $\mathcal{D}$ with parameters $\alpha_{\delta|i^c_{\mathrm{pa}(\delta)}} = (\alpha_{i_\delta|i^c_{\mathrm{pa}(\delta)}})_{i_\delta \in \mathcal{I}_\delta}$, also written as

$$(\theta_{\delta|i^c_{\mathrm{pa}(\delta)}}) \sim \mathcal{D}(\alpha_{\delta|i^c_{\mathrm{pa}(\delta)}}).$$

The probability density function for this Dirichlet distribution is given by

$$p(\theta_{\delta|i^c_{\mathrm{pa}(\delta)}}) \propto \prod_{i_\delta \in \mathcal{I}_\delta} (\theta_{i_\delta|i^c_{\mathrm{pa}(\delta)}})^{\alpha_{i_\delta|i^c_{\mathrm{pa}(\delta)}} - 1}.$$

By using Bayes' theorem, the posterior distribution is found to be

$$(\theta_{\delta|i^c_{\mathrm{pa}(\delta)}}|i^c) \sim \mathcal{D}(\alpha_{\delta|i^c_{\mathrm{pa}(\delta)}} + n_{\delta|i^c_{\mathrm{pa}(\delta)}}),$$

where the vector $n_{\delta|i^c_{\mathrm{pa}(\delta)}} = (n_{i_\delta|i_{\mathrm{pa}(\delta)}})_{i_\delta \in \mathcal{I}_\delta}$ contains zeros except at the place where $n_{i_\delta|i^c_{\mathrm{pa}(\delta)}} = n_{i^c_\delta|i^c_{\mathrm{pa}(\delta)}} = 1$. These numbers are also called counts as, when we update all the parameters recursively through the database d, $n_{i_\delta|i^c_{\mathrm{pa}(\delta)}}$ denotes the number of observations in d where δ and $pa(\delta)$ have that particular configuration.

4.3 Learning in the mixed case

In the mixed case we can write the local probability distributions as

$$
\begin{aligned}
&(Y_\gamma|i_{\mathrm{pa}(\gamma)}, y_{\mathrm{pa}(\gamma)}, \theta_{\gamma|i_{\mathrm{pa}(\gamma)}}) \sim \\
&\qquad\qquad \mathcal{N}(\beta^{+f}_{\gamma|i_{\mathrm{pa}(\gamma)}} z_{\mathrm{pa}(\gamma)}, \sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}),
\end{aligned}
$$

where

$$\beta^{+f}_{\gamma|i_{\mathrm{pa}(\gamma)}} = \begin{bmatrix} f_{\gamma|i_{\mathrm{pa}(\gamma)}} \\ \beta_{\gamma|i_{\mathrm{pa}(\gamma)}} \end{bmatrix} \quad \text{and} \quad z_{\mathrm{pa}(\gamma)} = \begin{bmatrix} 1 \\ y_{\mathrm{pa}(\gamma)} \end{bmatrix}$$

Notice that both these vectors have dimension $k+1$, where k is the number of continuous parents to γ.

As we assumed local independence for the discrete parents, we can, as in the discrete case, update the parameters for each configuration of the discrete parents independently. So consider a case $x^c \in d$ where the configuration of the discrete parents is $i^c_{\mathrm{pa}(\gamma)}$. In the following we do not use the index c on the parameters, as it will blur the notation.

A standard conjugate family for these observations is the family of Gaussian-inverse gamma distributions. Let the prior joint distribution of $\beta^{+f}_{\gamma|i_{\mathrm{pa}(\gamma)}}$ and $\sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}$ be as follows. The conditional prior distribution of $\beta^{+f}_{\gamma|i_{\mathrm{pa}(\gamma)}}$ given $\sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}$ is a multivariate Gaussian distribution and the marginal distribution of $\sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}$ is an inverse gamma distribution. The parameters are given as below.

$$(\beta^{+f}_{\gamma|i_{\mathrm{pa}(\gamma)}}|\sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}) \sim \mathcal{N}_{k+1}(\mu_{\gamma|i_{\mathrm{pa}(\gamma)}}, \sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}\tau^{-1}_{\gamma|i_{\mathrm{pa}(\gamma)}})$$
$$(\sigma^2_{\gamma|i_{\mathrm{pa}(\gamma)}}) \sim \mathcal{I}\Gamma\left(\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}, \frac{1}{\phi_{\gamma|i_{\mathrm{pa}(\gamma)}}}\right).$$

The parameters in the posterior distributions are easily found by Bayes' theorem, (DeGroot 1970).

5 Learning the structure of a CG network

Up until now we have assumed that the DAG D is known. In some situations this is not the case. Here we will show how we can select one or more DAG's among the possible DAG's. A way to find out how well a DAG represents the conditional independencies among the random variables in a Bayesian network, is to measure how likely the DAG is, given that we have observed a dataset d. That is, we can find the posterior probability of the DAG, $p(D|d)$. From Bayes' theorem we have that

$$p(D|d) \propto p(d|D)p(D).$$

As the normalizing constant does not depend upon structure, an often used measure, which gives the relative probability, is the network-score

$$p(D,d) = p(d|D)p(D).$$

In the next section we will derive the network-score for CG networks.

5.1 The network-score for a CG network

In order to calculate the network-score for a specific DAG D, we need to know the prior probability and the likelihood of the DAG. In this paper we do not consider how to find the prior probability of a DAG, but just note that we for example can let all DAG's be equally likely. The likelihood of the DAG D is given by

$$p(d|D) = \int_{\theta\in\Theta} p(d|\theta, D)p(\theta|D)d\theta, \qquad (7)$$

where Θ is the parameter space. Again we can consider the problem for the discrete part and the mixed part of the network separately. The discrete part is easily found to be

$$\prod_{\delta\in\Delta}\prod_{i_{\mathrm{pa}(\delta)}\in\mathcal{I}_{\mathrm{pa}(\delta)}} \frac{\Gamma(\alpha_{+\delta|i_{\mathrm{pa}(\delta)}})}{\Gamma(\alpha_{+\delta|i_{\mathrm{pa}(\delta)}} + n_{+\delta|i_{\mathrm{pa}(\delta)}})} \times \prod_{i_\delta\in\mathcal{I}_\delta} \frac{\Gamma(\alpha_{i_\delta|i_{\mathrm{pa}(\delta)}} + n_{i_\delta|i_{\mathrm{pa}(\delta)}})}{\Gamma(\alpha_{i_\delta|i_{\mathrm{pa}(\delta)}})}. \qquad (8)$$

In the mixed part of the network, the local marginal likelihoods are non-central t distributions with $\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}$ degrees of freedom, location vector $z_{\mathrm{pa}(\gamma)}\mu_{\gamma|i_{\mathrm{pa}(\gamma)}}$ and scale parameter $s_{\gamma|i_{\mathrm{pa}(\gamma)}} = \frac{\phi_{\gamma|i_{\mathrm{pa}(\gamma)}}}{\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}}(1 + (z_{\mathrm{pa}(\gamma)})^t\tau^{-1}_{\gamma|i_{\mathrm{pa}(\gamma)}}z_{\mathrm{pa}(\gamma)})$, see e.g. DeGroot (1970). So the mixed part is given by

$$\prod_{\gamma\in\Gamma}\prod_{i_{\mathrm{pa}(\gamma)}\in\mathcal{I}_{\mathrm{pa}(\gamma)}}\prod_{x^c\in d} \frac{\Gamma((\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}+1)/2)}{\Gamma(\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}/2)(\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}s_{\gamma|i_{\mathrm{pa}(\gamma)}}\pi)^{\frac{1}{2}}} \times (1+Q)^{\frac{(\rho_{\gamma|i_{\mathrm{pa}(\gamma)}}+1)}{2}}, \qquad (9)$$

where

$$Q = \frac{(y^c_\gamma - z^c_{\mathrm{pa}(\gamma)}\mu_{\gamma|i_{\mathrm{pa}(\gamma)}})^2}{\phi_{\gamma|i_{\mathrm{pa}(\gamma)}}(1 + (z^c_{\mathrm{pa}(\gamma)})^t\tau^{-1}_{\gamma|i_{\mathrm{pa}(\gamma)}}z^c_{\mathrm{pa}(\gamma)})},$$

and Γ is the gamma function. The network-score for a CG network is thus the product of the prior probability for the DAG D and the terms in (8) and (9). Notice that the network-score has the property that it factorizes into a product over terms involving only a single node and its parents. This property is called decomposability. So the network-score for CG networks is decomposable.

6 The master prior procedure

In the previous section we derived an expression for the network-score for CG networks. To calculate this score, we must specifying the local probability distributions and the local prior distributions for the parameters for each network under evaluation. In the papers Heckerman et al. (1995) and Geiger & Heckerman (1994) a method for finding the prior distributions for the parameters in respectively the pure discrete and the pure Gaussian case is developed. The work is based on principles of likelihood equivalence, parameter modularity, and parameter independence. It leads to a method where the parameter priors for all possible networks are deduced from one joint prior distribution, in the following called a master prior distribution. In this paper we will build on their method for finding a method, which can be used on networks with mixed variables. We will therefore in the following describe their method for the pure cases.

6.1 The master prior in the discrete case

In the pure discrete case, or the discrete part of a mixed network, the following is a well known classical result.

Let A be a subset of Δ and let $B = \Delta \setminus A$. Let the discrete variables i have the joint distribution

$$p(i|\Psi) = \Psi_i.$$

Notice here, that the set $\Psi = (\Psi_i)_{i\in\mathcal{I}}$ contains the parameters for the joint distribution, contrary to θ in Section 3, which contains the parameters for the conditional local distributions.

In the following let $z_{i_A} = \sum_{j:j_A=i_A} z_j$, where z is any parameter. Then the marginal distribution of i_A is given by

$$p(i_A|\Psi) = \Psi_{i_A},$$

and the conditional distribution of i_B given i_A is

$$p(i_B|i_A, \Psi) = \frac{\Psi_i}{\Psi_{i_A}} = \Psi_{i_A|i_B}$$

Further if the joint prior distribution for the parameters Ψ is Dirichlet, that is

$$p(\Psi) \sim \mathcal{D}(\alpha), \tag{10}$$

where $\alpha = (\alpha_i)_{i\in\mathcal{I}}$, then the marginal distribution of Ψ_A is Dirichlet, i.e.

$$p(\Psi_A) \sim \mathcal{D}(\alpha_A),$$

with $\alpha_A = (\alpha_{i_A})_{i_A\in\mathcal{I}_A}$. The conditional distribution of $\Psi_{B|i_A}$ is

$$p(\Psi_{B|i_A}) \sim \mathcal{D}(\alpha_{B|i_A})$$

with $\alpha_{i_A|i_B} = \alpha_i$. Furthermore the parameters are independent, that is

$$p(\Psi) = \prod_{i_A\in\mathcal{I}_A} p(\Psi_{B|i_A}) p(\Psi_A). \tag{11}$$

From the above result we see, that for each possible parent/child relationship, we can find the marginal parameter prior $p(\Psi_{\delta\cup\mathrm{pa}(\delta)})$. Further, from this marginal distribution we can, for each configuration of the parents, find the conditional local prior distribution $p(\Psi_{\delta|i_{\mathrm{pa}(\delta)}})$. Notice that $\Psi_{\delta|i_{\mathrm{pa}(\delta)}} = \theta_{\delta|i_{\mathrm{pa}(\delta)}}$, where $\theta_{\delta|i_{\mathrm{pa}(\delta)}}$ was specified for the conditional distributions in Section (3.1). Further, because of parameter independence, given by (11), we can find the joint parameter prior for any network as the product of the local priors involved.

To use this method, we must therefore specify the joint Dirichlet distribution, i.e. the master Dirichlet prior.

6.1.1 The master Dirichlet prior

We will now show how to construct the master Dirichlet prior. This was first done in Heckerman et al. (1995) and here we follow their method. We start by specifying a prior Bayesian network $(D, \mathcal{P})$ as we believe it to be. From this we calculate the joint distribution $p(i|\Psi) = \Psi_i$. As can be seen from (10), to specify a master Dirichlet distribution, we must specify the parameters $\alpha = (\alpha_{i_\delta})_{i\in\mathcal{I}}$. Consider now the following relation for the Dirichlet distribution.

$$p(i) = \mathbb{E}(\Psi_i) = \frac{\alpha_i}{n},$$

with $n = \sum_{i\in\mathcal{I}} \alpha_i$. Now we use the probabilities in the prior network as an estimate of $\mathbb{E}(\Psi_i)$, so we only need to determine n in order to calculate the parameters α_i. We determine n by using the notion of an imaginary database. We imagine that we have a database of cases, from which we from total ignorance have updated the distribution of Ψ. The sample size of this imaginary database is thus n. Therefore we refer to the estimate of n as the imaginary sample size, and it expresses how much confidence we have in the prior network.

6.2 The master prior in the Gaussian case

We have a similar result for the Gaussian case. Let A be a subset of Γ and let $B = \Gamma \setminus A$. If

$$(y|m, \Sigma) \sim \mathcal{N}(m, \Sigma),$$

then

$$(y_A|m, \Sigma) \sim \mathcal{N}(m_A, \Sigma_{AA})$$

and

$$(y_B|y_A, m_{B|A}, \beta_{B|A}, \Sigma_{B|A}) \sim \mathcal{N}(m_{B|A} + \beta_{B|A} y_A, \Sigma_{B|A}),$$

where

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}, \; \Sigma_{B|A} = \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB},$$

$$m_{B|A} = m_B - \beta_{B|A} m_A \quad \text{and} \quad \beta_{B|A} = \Sigma_{BA}\Sigma_{AA}^{-1}.$$

Further, if

$$(m|\Sigma) \sim \mathcal{N}(\mu, \frac{1}{\nu}\Sigma) \;\text{ and }\; (\Sigma) \sim \mathcal{IW}(\rho, \Phi),$$

where the parametric matrix Φ is partitioned as Σ, then

- $(\Sigma_{AA}) \sim \mathcal{IW}(\rho, \Phi_{AA})$
- $(\Sigma_{B|A}) \sim \mathcal{IW}(\rho + |A|, \Phi_{B|A})$
- $(m_{B|A}, \beta_{B|A}|\Sigma_{B|A}) \sim \mathcal{N}(\mu_{B|A}, \Sigma_{B|A} \otimes \tau_{B|A}^{-1})$
- $m_A, \Sigma_{AA} \perp\!\!\!\perp m_{B|A}, \beta_{B|A}\Sigma_{B|A}$

where

$$\mu_{B|A} = (\mu_B - \Phi_{BA}\Phi_{AA}^{-1}\mu_A, \Phi_{BA}\Phi_{AA}^{-1}),$$

and

$$\tau_{B|A}^{-1} = \begin{pmatrix} \frac{1}{\nu} & -\mu_A^T \Phi_{AA}^{-1} \\ \Phi_{AA}^{-1}\mu_A & \Phi_{AA}^{-1} \end{pmatrix},$$

and $\otimes$ denotes the Kronecker product. Notice that the dimension of $\mu_{B|A}$ is $(|B|, |B| \times |A|)$.

As in the discrete case, this result shows us how to deduce the local probability distributions and the local prior distributions from the joint distributions. Further we can, again because of parameter independence, specify the joint parameter prior for any Gaussian network as the product of the local priors. Notice again that the parameters found here for a node given its parents, coincides with the parameters specified in Section 3.2.

6.2.1 The master Gaussian-inverse Wishart prior

Before we show how to construct the master prior, we need the following result. The Gaussian-inverse Wishart prior is conjugate to observations from a Gaussian distribution, (DeGroot 1970). So let the probability distribution and the prior distribution be given as above. Then, given the database $d = \{y^1, \ldots, y^n\}$, the posterior distributions are

$$(m|\Sigma, d) \sim \mathcal{N}(\mu', \frac{1}{\nu'}\Sigma) \text{ and } (\Sigma|d) \sim \mathcal{IW}(\rho', \Phi'),$$

where

$$\begin{aligned} \nu' &= \nu + n \\ \mu' &= \frac{\nu\mu + n\overline{y}}{\nu + n} \\ \rho' &= \rho + n \\ \Phi' &= \Phi + ssd + \frac{\nu n}{\nu + n}(\mu - \overline{y})(\mu - \overline{y})^t, \end{aligned} \tag{12}$$

with

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad ssd = \sum_{i=1}^{n}(y_i - \overline{y})(y_i - \overline{y})$$

From these updating formulas we see that ν' and ρ' are updated with the number of cases in the database. Further μ' is a weighted average of the prior mean and the sample mean, each weighted by their sample sizes. Finally Φ is updated with the ssd, which expresses how much each observation differs from the sample mean, and an expression for how much the prior mean differs from the sample mean.

To specify the master prior, we need to specify the four parameters ν, μ, ρ and Φ. As for the discrete variables we start by specifying a prior Bayesian network, $(D, \mathcal{P})$. From this we can deduce a prior joint probability distribution $p(y|m, \Sigma) = \mathcal{N}(m, \Sigma)$. We now imagine that the mean m and the variance Σ were calculated from an imaginary database, so that they actually are the sample mean and the sample variance. Further we assume that before we observed this imaginary database, we were totally ignorant about the parameters. We can now use the formulas in (12) to "update" the parameters on the basis of our imaginary database. As we have not seen any cases before, ν and ρ are estimated by the imaginary sample size. Further

$$\mu = m \quad \text{and} \quad \Phi = ssd = (\nu - 1)\Sigma.$$

In Geiger & Heckerman (1994), μ and Φ are found in a slightly different way.

6.3 Properties of the master prior procedure

The method for finding prior parameter distributions described in the previous section, has some properties, which we will describe here. In the following we use Ψ as parameters defined for joint distribution, i.e. Ψ can be the parameter for the discrete variables or in the continuous case, $\Psi = (m, \Sigma)$.

Clearly a consequence of using the method is that the parameters are independent. Further it can be seen, that if a node v has the same parents in two DAG's D_1 and D_2, then

$$p(\Psi_{v|\mathrm{pa}(v)}|D_1) = p(\Psi_{v|\mathrm{pa}(v)}|D_2)$$

This property is referred to as parameter modularity. Now the discrete and Gaussian distributions have the property that if the joint probability distribution $p(x)$ can be factorized according to a DAG D, then it can also be factorized according to all other DAG's, which represents the same set of condtional independencies as D. A set of DAG's, D^e, which represents the same independence constraints is referred to as independence equivalent DAG's. So let D_1 and D_2 be independence equivalent DAG's, then

$$p(x|\Psi, D_1) = p(x|\Psi, D_2).$$

This means, that from observations alone we can not distinguish between different DAG's in an equivalence class. In the papers Heckerman et al. (1995) and Geiger & Heckerman (1994) it is for respectively the discrete and Gaussian cases shown, that when using the master prior procedure for construction parameter priors, the marginal likelihood for data is also the same for independence equivalent networks, i.e.

$$p(d|D_1) = p(d|D_2)$$

This equivalence is referred to as likelihood equivalence. Note that likelihood equivalence imply, that if D_1 and D_2 are independence equivalent networks, then they have the same joint prior for the parameters, i.e. $p(\Psi|D_1) = p(\Psi|D_2)$.

7 Local masters for mixed networks

In this section we will show how to specify prior distributions for the parameters in a CG network. In the mixed case, the marginal of a CG distribution is not always a CG distribution. In fact it is only a CG distribution if we marginalize over continuous variables or if we marginalize over a set B of discrete variable, where $B \perp\!\!\!\perp \Gamma \mid \Delta \setminus B$, see Frydenberg (1990). Consider the following example. We have a network of two variables i and y and the joint distribution is given by

$$p(i, y) = p(i)\mathcal{N}(m_i, \sigma_i^2)$$

Then the marginal distribution of y is given as a mixture of normal distributions

$$p(y) = \sum_{i \in \mathcal{I}} p(i)\mathcal{N}(m_i, \sigma_i^2),$$

so there is no simple way of using this directly for finding the local priors.

7.1 The suggested solution

The suggested solution is very similar to the solution for the pure cases. We start by specifying a prior Bayesian network $(D, \mathcal{P})$ and then calculate the joint probability distribution

$$p(i, y|H) = p(i|\Psi)\mathcal{N}(m_i, \Sigma_i),$$

with $H = (\Psi, (m_i)_{i\in\mathcal{I}}, (\Sigma_i)_{i\in\mathcal{I}})$, i.e. from the conditional parameters in the local distributions in the prior network, we calculate the parameters for the joint distribution. Then we translate this prior network into an imaginary database, with imaginary sample size n, where n depends on how certain we are of the prior network. From the probabilities in the discrete part of the network, we can, as in the pure discrete case, calculate α_i for all configurations of i. Now α_i represents how many observation of $I = i$ we have in the imaginary database. We assume, that each time we have observed the discrete variables I, we have observed the continuous variables Y and therefore we set $\nu_i = \rho_i = \alpha_i$. Now for each configuration of i we let m_i be the sample mean in the imaginary database, and Σ_i the sample variance. Further, as for the pure Gaussian case, we use $m_i = \mu_i$ and $\Phi_i = (\nu_i - 1)\Sigma_i$. We have now specified all the parameters needed to define the joint prior distributions for the parameters, so

$$\begin{aligned} p(\Psi) &= \mathcal{D}(\alpha) \\ p(m_i|\Sigma_i) &= \mathcal{N}(\mu_i, \frac{1}{\nu_i}\Sigma_i) \\ p(\Sigma_i) &= \mathcal{IW}(\rho_i, \Phi_i), \end{aligned}$$

But we can not use these distributions to derive priors for other networks, so instead we use the imaginary database to derive local master distributions.

Let for each family $A = v \cup \mathrm{pa}(v)$ the marginal probability distribution be given by

$$p(x_A|H_A) = CG(\Psi_{i_{A\cap\Delta}}, (m_{i_{A\cap\Delta}})_{A\cap\Gamma}, (\Sigma_{i_{A\cap\Delta}})_{A\cap\Gamma}).$$

Then we suggest that the marginal prior distributions, also called the local masters, are found in the following way:

Let $z_{i_{A\cap\Delta}} = \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} z_j$. Then

$$\begin{aligned} (\Psi_{A\cap\Delta}) &\sim \mathcal{D}(\alpha_{A\cap\Delta}) \\ ((\Sigma_{i_{A\cap\Delta}})_{A\cap\Gamma}) &\sim \mathcal{IW}(\rho_{i_{A\cap\Delta}}, (\tilde{\Phi}_{i_{A\cap\Delta}})_{A\cap\Gamma}) \end{aligned}$$

and

$$((m_{i_{A\cap\Delta}})_{A\cap\Gamma}|(\Sigma_{i_{A\cap\Delta}})_{A\cap\Gamma}) \sim \mathcal{N}((\overline{\mu}_{i_{A\cap\Delta}})_{A\cap\Gamma}, \frac{1}{\nu_{i_{A\cap\Delta}}}(\Sigma_{i_{A\cap\Delta}})_{A\cap\Gamma}),$$

where

$$\overline{\mu}_{i_{A\cap\Delta}} = \frac{(\sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \mu_j\nu_j)}{\nu_{i_{A\cap\Delta}}},$$

and

$$\tilde{\Phi}_{i_{A\cap\Delta}} = \Phi_{i_{A\cap\Delta}} + \sum_{j:j_{A\cap\Delta}=i_{A\cap\Delta}} \nu_j(\mu_j - \overline{\mu}_{i_{A\cap\Delta}})(\mu_j - \overline{\mu}_{i_{A\cap\Delta}})^t$$

The equations in the above result is well known in the analysis of variance theory. The marginal mean is found as a weighted average of the mean in every group, where a group here is given as a configuration of the discrete parents we marginalize over. The weights are the number of observations in each group. The marginal *ssd* is given as the within group variation plus the between group variation. Notice that with this method it is possible to specify mixed networks, where the mean in the mixed part of the network does not depend on the discrete parents, but the variance does (and vice versa).

From the local masters we can now, by conditioning as in the pure cases, derive the local priors needed to specify the prior parameter distribution for a CG network. So the only difference between the master procedure and the local master procedure is in the way the marginal distributions are found.

7.2 Properties of the local master procedure

The local master procedure coincides with the master procedure in the pure cases. Further, the properties of the local master procedure in the mixed case, are the same as of the master prior procedure in the pure cases.

Parameter independence and parameter modularity follows immediately from the definition of the procedure. To show likelihood equivalence, we need the following result from Chickering (1995). Let D_1 and D_2 be two DAG's and let R_{D_1,D_2} be the set of edges by which D_1 and D_2 differ in directionality. Then, D_1 and D_2 are independence equivalent if and only if there exists an sequence of $|R_{D_1,D_2}|$ distinct arc reversals applied to D_1 with the following properties:

- After each reversal, the resulting network structure is a DAG, i.e. it contains no directed cycles and it is independence equivalent to D_2.
- After all reversals, the resulting DAG is identical to D_2.
- If $w \to v$ is the next arc to be reversed in the current DAG, then w and v have the same parents in both DAG's, with the exception that w is also a parent of v in D_1.

Note that as we only reverse $|R_{D_1,D_2}|$ distinct arcs, we only reverse arcs in R_{D_1,D_2}. For mixed networks this means that we only reverse arcs between discrete variables or between continuous variables, as the only arcs that can differ in directionality are these. So we can use the above result for mixed networks.

From the above we see, that we can show likelihood equivalence by showing that $p(d|D_1) = p(d|D_2)$ for two independence equivalent DAG's D_1 and D_2 that differ only by the direction of a single arc. As $p(x|H, D_1) = p(x|H, D_2)$ in CG networks, we can show likelihood equivalence by showing that $p(H|D_1) = p(H|D_2)$.

In the following let $v \to w$ in D_1 and $w \to v$ in D_2. Further let ∇ be the set of common discrete and continuous parents for v and w. Of course if v and w are discrete variables, then ∇ only contains discrete variables. The relation between $p(H|D_1)$ and $p(H|D_2)$ is given by:

$$\begin{aligned} \frac{p(H|D_1)}{p(H|D_2)} &= \frac{p(H_{v|w\cup\nabla}, D_1)p(H_{w|\nabla}, D_1)}{p(H_{w|v\cup\nabla}, D_2)p(H_{v|\nabla}, D_2)} \\ &= \frac{p(H_{v\cup w|\nabla}, D_1)}{p(H_{v\cup w|\nabla}, D_2)} \end{aligned} \tag{13}$$

When using the local Master procedure, the terms in (13) are equal. This is evident, as we find the conditional priors from distributions over families A, in this case $A = v \cup w \cup \nabla$, which is the same for both networks. Therefore likelihood equivalence follows.

Acknowledgements

This research was supported by the ESPRIT projekt P29105 (BaKE). Also I would like to thank my supervisor Steffen L. Lauritzen for his dedicated help and support.

References

Chickering, D. (1995). A transformational characterization of equivalent Bayesian-network structures, *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence* pp. 87–98. Montreal, QU.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models, *Annals of Statistics* **18**: 790–805.

Geiger, D. & Heckerman, D. (1994). Learning Gaussian Networks, *Technical Report MSR-TR-94-10*, Microsoft Research.

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

Lauritzen, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models, *Journal of the American Statistical Association* **87**(420): 1098–1108.

Lauritzen, S. L. (1996). *Graphical Models*, Clarendon press, Oxford, New York.

Lauritzen, S. L. & Jensen, F. (1999). Stable Local Computation with Conditional Gaussian Distributions, *Technical Report R-99-2014*, Aalborg University, Denmark.

Shachter, R. D. & Kenley, C. R. (1989). Gaussian influence diagrams, *Management Science* **35**: 527–550.

Spiegelhalter, D. J. & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures, *Networks* **20**: 579–605.