Accelerated Quantification of Bayesian Networks with Incomplete Data

Bo Thiesson
Aalborg University
Fredrik Bajers Vej 7E
DK-9220 Aalborg Ø, Denmark
thiesson@iesd.auc.dk

Abstract

Probabilistic expert systems based on Bayesian networks (BNs) require initial specification of both a qualitative graphical structure and quantitative assessment of conditional probability tables. This paper considers statistical batch learning of the probability tables on the basis of incomplete data and expert knowledge. The EM algorithm with a generalized conjugate gradient acceleration method has been dedicated to quantification of BNs by maximum posterior likelihood estimation for a super-class of the recursive graphical models. This new class of models allows a great variety of local functional restrictions to be imposed on the statistical model, which hereby extents the control and applicability of the constructed method for quantifying BNs.

Introduction

The construction of probabilistic expert systems (Pearl 1988, Andreassen *et al.* 1989) based on Bayesian networks (BNs) is often a challenging process. It is typically divided into two parts: First the construction of a graphical structure which defines relations between variables in a model, and second the quantitative assessment of the strength of these relations as defined by tables of conditional distributions.

Both aspects of this process can be eased by applying automated methods for learning from a database of observations or from a combination of a database and expert knowledge. See Buntine (1995) for a literature review on different learning methods.

This paper considers statistical batch learning for the quantitative assessment of relations in a given structural model. In this scheme a BN resembles a quantified statistical model, that is, a particular distribution belonging to the set of distributions as defined by the model. Usually, the recursive graphical models of Wermuth & Lauritzen (1983) underlie the statistical modelling for BNs. We investigate a superclass for these models, denoted as recursive exponential models (REMs), which have evolved by the desire to impose functional restrictions onto local components of the model. One may visualize a local component as the part of a model which defines the functional structure of a particular conditional distribution in a quantification table for the BN. Hence, the REMs extends the recursive graphical models by the possibility to functionally control the quantification of these tables.

Given a database of observations, the maximum likelihood estimate (MLE) is the usual candidate for a quantification of a model. If also prior knowledge on parameters is available, the largest posterior mode is a natural alternative. This mode will be denoted as the maximum *posterior* likelihood estimate (MPLE).

In situations of incomplete observations the determination of the MLE or MPLE may call for numerical techniques. In Lauritzen (1995) it is shown how to exploit the EM algorithm, as formally defined in Dempster *et al.* (1977), for maximization within the framework of recursive graphical models. Unlike various other numerical techniques for optimization, the EM algorithm converges reliably even when started in a distant point from the maximum. A common criticism, however, is that convergence can be painfully slow if high accuracy of the estimate is necessary. See e.g. Louis (1982), Meilijson (1989), Jamshidian & Jennrich (1993), and several discussants of Dempster *et al.* (1977). For this reason we investigate an acceleration of the EM algorithm by a generalised conjugate gradient algorithm which has a superior rate of convergence when started close enough to the maximum to ensure convergence. By using the EM algorithm for early iterations and the acceleration algorithm for the final iterations, we have constructed a hybrid method which preserves the global convergence of the EM algorithm but has a higher rate of convergence.

The idea of accelerating the EM algorithm is not

[0] The author is grateful to Steffen L. Lauritzen for helpful discussions. Grant support was provided by ESPRIT Basic Research Action, project no. 6156 (DRUMS II) and the PIFT programme of the Danish Research Councils.

new in general. See Jamshidian & Jennrich (1993) for a recent survey. In the context of quantifying BNs it is new, probably due to the lack of publications on the analytical evaluation of derivatives as needed for most accelerations. Lauritzen (1995) and Spiegelhalter *et al.* (1993) do, however, mention the possibility of calculating the gradient for recursive graphical models, and in fact Russell *et al.* (1995) covers gradient-descent methods for MLE quantification based on these models.

Section 2 and 3 review the EM algorithm and the generalized conjugate gradient algorithm used for acceleration. Section 4 gives a concise description of REMs and account for the simplification into the well-known recursive graphical models. In Section 5 the algorithms are specialized for these models.

The MLE method

Given a conceptual model, yielding the vector of random variables $X = (X_v)_{v \in V}$, with a family of distributions $p(X|\theta)$ parameterised by the vector $\theta \in \Theta$ and denote by $l(\theta|x)$, the associated log-likelihood function.

Suppose that x is only observed indirectly through the actually observed, possible incomplete, sample of data y. The observed data may be incomplete in two ways: Observations may be *missing* according to a full instantiation of X so that individual cases only hold observed values according to a sub-vector $X_A, A \subset V$. This accounts for both situations of latent (or hidden) variables and situations of randomly missing values among variables. Observations may also be *imprecise* if there are variables for which the collector of data cannot distinguish between a set of possible values and therefore reports this set instead of just a single value.

By this scheme, incomplete data associates a set of possible completions denoted $\chi(y)$. Under the condition that the observed data is incomplete in an uninformative way according to these possible completions, the incomplete data distribution satisfies

$$p(y|\theta) = \sum_{x \in \chi(y)} p(x|\theta). \tag{1}$$

In case of incomplete data the MLE is typically too difficult to calculate analytically. Here, the general idea of the numerical approaches considered in this paper is described.

The EM algorithm

The EM algorithm has an intuitively easy interpretation of converting the ML estimation into a sequence of "pseudo-estimations" with respect to the conceptual model for complete data. Let θ^n denote the current value of θ after n iterations. Each iteration of the EM algorithm can then be described in two steps:

E-step: Determine the conditional expectation of the log-likelihood function given the observed data

$$Q(\theta|\theta^n, y) = \mathbb{E}_{\theta^n}[l(\theta|X)|y].$$

M-step: Determine θ^{n+1} by maximizing $Q(\theta|\theta^n, y)$ in θ.

Generalizations of the EM algorithm appear by strictly increasing $Q(\theta^{n+1}|\theta^n, y)$ over $Q(\theta^n|\theta^n, y)$ rather than maximising it. The generalized EM algorithms may be favourable in situations where the MLE of $Q(\theta|\theta^n, y)$ has to be calculated iteratively.

There is no clear criterion for when to "retire" the EM algorithm in favour of the acceleration algorithm. However, if acceleration is started too early, divergence will reveal by a sudden decrease in likelihood value. If this happens the EM algorithm must take over a few iterations from the point previous to the decrease in likelihood before the faster method is started again.

Conjugate gradient acceleration

On the ground that appropriate generalised gradients can significantly improve the performance of algorithms that use gradients, Jamshidian & Jennrich (1993) proposed an acceleration of the EM algorithm based on a generalised conjugate gradient method.

They showed that if the MLE $\hat{\theta}$ is an interior point of Θ, then

$$\theta^{n+1} - \theta^n = -(\ddot{Q}(\hat{\theta}|\hat{\theta}, y))^{-1}\dot{l}(\theta^n|y) + o(\theta^n - \hat{\theta}),$$

where $\dot{l}(\theta^n|y)$ is the gradient of the log-likelihood function at θ^n and $\ddot{Q}(\hat{\theta}|\hat{\theta}, y))$ is the Hessian of $Q(\theta|\hat{\theta}, y)$ evaluated at $\hat{\theta}$. As the key to the acceleration method they observed that in the neighbourhood of $\hat{\theta}$, the EM-step $\theta^{n+1} - \theta^n$ approximates the generalised gradient $\tilde{l}(\theta^n|y) = -(\ddot{Q}(\hat{\theta}|\hat{\theta}, y))^{-1}\dot{l}(\theta^n|y)$ with the norm $\| \theta \| = (\theta'(-\ddot{Q}(\hat{\theta}|\hat{\theta}, y))\theta)^{\frac{1}{2}}$, where $'$ denotes transpose.

The obvious advantage of this approximation is that the evaluation of a generalized gradient does only require an EM-step. Hence, by the assumption that the EM-step qualifies as an appropriate generalised gradient, Jamshidian & Jennrich (1993) proposed a generalised conjugate gradient acceleration, which for each iteration operates as follows:

LS-step (Line search): Determine $\theta^{n+1} = \theta^n + \alpha d_n$, where α is a step length which (approximately) maximizes $l(\theta^n + \alpha d_n|y)$.

DU-step (Direction update): Determine the next conjugate direction as

$$\begin{aligned} d_{n+1} &= \tilde{l}(\theta^{n+1}|y) - \beta d_n, \text{where} \\ \beta &= \frac{\tilde{l}(\theta^{n+1}|y)'(\dot{l}(\theta^{n+1}|y) - \dot{l}(\theta^n|y))}{d_n'(\dot{l}(\theta^{n+1}|y) - \dot{l}(\theta^n|y))}. \end{aligned}$$

The algorithm is initialised by $d_0 = \tilde{l}(\theta^0|y)$.

The algorithm is motivated as an acceleration of the EM algorithm as follows. If the length of an EM-step is not optimal, a line search in this direction will improve the rate of convergence. This is actually the acceleration algorithm with $\beta \equiv 0$. The rate of convergence can also be improved by ensuring that moving along in a direction will not cancel out traversals of previous steps. Hence, instead of moving in the direction of the EM-step, we proceed in a direction which is conjugate to the previous direction, and, insofar as possible, to all previous directions traversed. This is accomplished by β with the evaluation of gradients as the cost.

The MPLE method

Traditional ML estimation may be compromised when dealing with ill-posed problems like latent structure models or situations of sparse data, where small changes in the data can have a large impact on the estimate. In these situations we may resort to a Bayesian interpretation given to the estimation problem. Instead of just maximizing the likelihood we may incorporate prior information about the parameters by finding the largest posterior mode, the MPLE. Dempster *et al.* (1977) briefly describe how to modify the EM algorithm to produce the MPLE. This has been further entertained in Green (1990). A specialization for recursive graphical models is found in Lauritzen (1995).

Suppose we have information about θ in the form of a prior distribution $\pi(\theta)$, then

$$p(\theta|y) \propto p(y|\theta)\pi(\theta).$$

By considering the posterior distribution as a posterior likelihood, maximized by the EM algorithm by simply replacing the E-step with the expectation of the posterior log-likelihood, $Q^*(\theta|\theta^n)$, the E-step becomes

$$Q^*(\theta|\theta^n) = Q(\theta|\theta^n) + \log \pi(\theta). \qquad (2)$$

In the M-step, Q^* is maximized instead of Q.

The Bayesian interpretation of the EM algorithm can be projected directly onto the gradient, as additionally needed for the acceleration method.

Analogous to the notation of the gradient for the (traditional) log-likelihood by $\dot{l}(\theta|y)$, let $\dot{l}(\theta)$ and $\dot{l}^*(\theta|y)$ denote the gradients for the logarithm of prior and posterior distributions, respectively. The gradient of the posterior log-likelihood is then given by

$$\dot{l}^*(\theta|y) = \dot{l}(\theta|y) + \dot{l}(\theta). \qquad (3)$$

In effect, each of the expressions which goes into the MPLE method is made up by two terms, which describe the game between fidelity and amount of data against prior knowledge for modelling an acceptable solution to the problem.

The statistical modelling

Here, we introduce the REMs, which have evolved from the recursive graphical models by the desire to impose functional restrictions onto local components of a model. We also investigate appropriate priors.

Recursive exponential models

A REM can be graphically represented by a directed acyclic graph. That is, the variables X can be arranged by a response structure, where the set of nodes represents variables and directed edges signify for each variable $X_v \in X$ the existence of direct causal influence from variables represented by the parents $X_{pa(v)}$.

According to this graphical structure, a REM holds assumptions of variation independence between parameters in different local components of the model, to be described below. Readers familiar with Spiegelhalter & Lauritzen (1990) and the line of work reported in Heckerman *et al.* (1995) may recognize these assumptions as used in these papers but not explicitly named.

By *global variation independence*, $p(X|\theta)$ factorises into a product of mutually independent components given by the recursive response structure of the graph. That is,

$$p(X|\theta) = \prod_{v \in V} p(X_v|X_{pa(v)}, \theta_v),$$

where $\Theta = \times_{v \in V} \Theta_v$, and $\theta_v \in \Theta_v$ completely specifies the relationship between the variable X_v and its conditional set of variables $X_{pa(v)}$.

In some applications, particularly pedigree analysis, it is typical to restrain the tables of conditional distributions by using the knowledge that some of the tables are equal. Equal tables must be parametrized by the same parameters. Let $\tilde{v} \subseteq V$ specify a set of variables that associates equal tables and denote by $\tilde{V}$ the total set of these equivalence classes. Then

$$p(X|\theta) = \prod_{\tilde{v} \in \tilde{V}} \prod_{v \in \tilde{v}} p(X_v|X_{pa(v)}, \theta_{\tilde{v}}),$$

where $\theta_{\tilde{v}} \in \Theta_{\tilde{v}}$ specifies the relationship between X_v and $X_{pa(v)}$ for any $v \in \tilde{v}$. Hence, the global parameter independence is relaxed a bit. If equal tables are represented by a single generic table, as assumed from now on, this representation more directly illustrates the reduction of the parameter space into $\Theta = \times_{\tilde{v} \in \tilde{V}} \Theta_{\tilde{v}}$.

By *local variation independence* each (generic) table is additionally broken into local components of conditional distributions as defined for each parent configuration. For $v \in \tilde{v}$, let $r = 1, \ldots, R_{\tilde{v}}$ index a parent configuration in the generic table, and let $s = 0, \ldots, S_{\tilde{v}}$ index a particular value of the variable.

Hence, $\Theta_{\bar{v}} = \times_r \Theta_{\bar{v}|r}$ and conditional probabilities are given by $p(s|r, \theta_{\bar{v}|r})$, where $\theta_{\bar{v}|r} \in \Theta_{\bar{v}|r}$.

By these simplifying assumptions the quantification of a REM is broken into the quantification of local models, which comply with the typical scenario of breaking down the quantification of a BN into tables of independently quantified conditional distributions.

The statistical modelling by REMs does not stop at this point. To completely qualify as a REM each local model must be structurally defined by a regular exponential model. As any exponential model is allowed, the REMs become a very extensive class of models, which allows sophisticated functional restrictions to be placed on each local model, if necessary.

Disregarding the possibility of specifying equal tables, the local exponential modelling makes the difference from the recursive graphical models for which each local model cannot be restricted beyond the fact that it is a model of probability distributions. We do not account this as a functional restriction.

A recursive graphical model is defined in the framework of REMs as follows (positivety constraints are applied for simplicity). Consider the local model that structurally defines the conditional distribution $p(\cdot|r)$. Let s_0 denote an index of reference, say $s_0 = 0$, and let for $s_+ = 1, \ldots, S_{\bar{v}}$

$$\theta^{s_+} = \log[p(s_+|r)/p(s_0|r)]$$

and

$$t^{s_+}(s) = \begin{cases} 1 \text{ for } s = s_+ \\ 0 \text{ otherwise.} \end{cases}$$

The local model is then defined by the exponential model having probabilities of the form

$$p(s|r, \theta) = b(s, r) \exp[\theta' t(s) - \phi(\theta)], \qquad (4)$$

$$\phi(\theta) = \log \left(\sum_{s_+=1}^{S_{\bar{v}}} \exp(\theta^{s_+}) \right),$$

where $\theta = (\theta^1, \ldots, \theta^{S_{\bar{v}}})$ defines the parameters, $t(s) = (t^1(s), \ldots, t^{S_{\bar{v}}}(s))$ the statistics, $\phi(\theta)$ the normalizing function, and $b(s, r) = 1$ the carrying density.

Prior distributions for parameters

The construction of a prior distribution for parameters is simplified considerably by matching the assumptions of variation independence with assumptions of relaxed *global and local independence* of parameters considered as random variables. By these assumptions, the distribution for parameters factorises as

$$\pi(\theta) = \prod_{\bar{v} \in \bar{V}} \prod_{r=1}^{R_{\bar{v}}} \pi(\theta_{\bar{v}|r}).$$

Hence, each local component of the prior can be considered independently.

The notion of global and local independence is also nicely covered within the line of work reported in Heckerman *et al.* (1995). It is inspired by similar assumptions in Spiegelhalter & Lauritzen (1990), which introduced a method for sequential updating a Bayesian network as new observations eventuate. To prepare the quantification methods for the possibility of future sequential updating by this method, we are especially interested in (approximately) conjugate priors.

If functional restrictions are not specified for the local model, the natural conjugate prior on probabilities is given by a $S_{\bar{v}}$-dimensional Dirichlet distribution with parameters $\alpha(s, r)$ associated for each probability. That is, $p(\cdot|r, \theta_{\bar{v}|r}) \sim \mathcal{D}(\alpha(0, r), \ldots, \alpha(S_{\bar{v}}, r))$. By a transformation of parameters as given by the exponential representation of probabilities in (4), the prior distribution for $\theta_{\bar{v}|r}$ is defined by (noting that $dp(\cdot|r, \theta_{\bar{v}|r})/d\theta_{\bar{v}|r} = \prod_{s=0}^{S_{\bar{v}}} p(s|r, \theta_{\bar{v}|r})$ gives the Jacobian of the transformation)

$$\pi(\theta_{\bar{v}|r}) \propto \prod_{s=0}^{S_{\bar{v}}} p(s|r, \theta_{\bar{v}|r})^{\alpha(s,r)}. \qquad (5)$$

Given a general exponential local model, the construction of a conjugate prior becomes more complicated. Denote by θ^* the value that maximizes the local prior $\pi(\theta_{\bar{v}|r})$. By a Taylor series expansion around θ^* Thiesson (1995) shows that a conjugate distribution can be approximated by a distribution proportional to the multivariate normal distribution

$$\mathcal{N}(\theta^*, \frac{1}{\beta} I(\theta^*)^{-1}), \qquad (6)$$

where β and the maximizing value θ^* are unknown parameters to be assessed by experts, and $I(\theta^*)$ denotes the observed information at the value θ^*.

In practice though, it seems unreasonable to request domain experts for a parametrization of any of these priors. To overcome this problem Thiesson (1995) also shows how to assess the parametrization from a specification of a "best guess" distribution with a judgment of imprecision (or confidence) on each of the probabilities in the form of an upper and lower boundary. Assessment of Dirichlet priors can also be studied in Spiegelhalter *et al.* (1993) and Heckerman *et al.* (1995).

Specialization

The maximization algorithms are specialized for the REMs. We consider computation of the MPLE, but the MLE is easily obtained by inserting non-informative priors in the following. It turns out that maximization can be accomplished by local computations.

The EM algorithm

To identify the E-step, we will first consider the likelihood function for a sample of independent complete observations, $x = (x^1, \ldots, x^L)$. Due to the factorization of the probability for a single observation

$$p(x^l|\theta) = \prod_{\tilde{v}\in\tilde{V}} \prod_{v\in\tilde{v}} p(x_v^l|x_{pa(v)}^l, \theta_{\tilde{v}|x_{pa(v)}^l})$$

the likelihood factorises as

$$L(\theta|x) \propto \prod_{l=1}^{L} p(x^l|\theta) = \prod_{\tilde{v}\in\tilde{V}} \prod_{r=1}^{R_{\tilde{v}}} \prod_{s=0}^{S_{\tilde{v}}} p(s|r, \theta_{\tilde{v}|r})^{\tilde{n}(s,r)},$$

where $\tilde{n}(s,r) = \sum_{v\in\tilde{v}} n(s,r)$, and $n(s,r)$ denotes the marginal count in the configuration (s,r) for a family of variables $(X_v, X_{pa(v)})$. The marginal count is obtained by adding up the qualifying observations as

$$n(s,r) = \sum_{l=1}^{L} \chi^{(s,r)}(x_v^l, x_{pa(v)}^l)$$

where

$$\chi^{(s,r)}(x_v^l, x_{pa(v)}^l) = \begin{cases} 1 \text{ for } (x_v^l, x_{pa(v)}^l) = (s,r) \\ 0 \text{ otherwise.} \end{cases} \qquad (7)$$

For a sample of independent, possibly incomplete, observations $y = (y^1, \ldots, y^L)$ the conditional expectation of the likelihood function is obtained by replacing the marginal counts by expected marginal counts

$$\begin{aligned} \tilde{n}^*(s,r) &= \sum_{v\in\tilde{v}} n^*(s,r) \\ &= \sum_{v\in\tilde{v}} \sum_{l=1}^{L} p(X_v{=}s, X_{pa(v)}{=}r|y^l, \theta). \quad (8) \end{aligned}$$

As pointed out in Lauritzen (1995), the posterior probabilities in (8) can be efficiently calculated by the procedure of Lauritzen & Spiegelhalter (1988) for probability propagation.

The E-step (2) can now be identified as

$$\begin{aligned} &Q^*(\theta|\theta^n, y) \\ &= \sum_{\tilde{v}\in\tilde{V}} \sum_{r=1}^{R_{\tilde{v}}} \left[\sum_{s=0}^{S_{\tilde{v}}} \tilde{n}^*(s,r) \log p(s|r, \theta_{\tilde{v}|r}) + \log \pi(\theta_{\tilde{v}|r}) \right] \\ &= \sum_{\tilde{v}\in\tilde{V}} \sum_{r=1}^{R_{\tilde{v}}} Q^*(\theta_{\tilde{v}|r}|\theta^n, y), \end{aligned}$$

where $p(s|r, \theta_{\tilde{v}|r})$ is of the exponential form (4).

By this, the M-step is completed by maximizing (or increasing) each local part of the expected log-likelihood independently.

If the local model does not hold functional restrictions, the local prior is given by (5), and the maximum for $Q^*(\theta_{\tilde{v}|r}|\theta^n, y)$ can be found analytically as the value $\hat{\theta}_{\tilde{v}|r} \in \Theta_{\tilde{v}|r}$ which obeys

$$p(s|r, \hat{\theta}_{\tilde{v}|r}) = \frac{\tilde{n}^*(s,r) + \alpha(s,r)}{\tilde{n}^*(r) + \alpha(r)},$$

where $\tilde{n}^*(r) = \sum_{s=0}^{S_{\tilde{v}}} \tilde{n}^*(s,r)$ and $\alpha(r) = \sum_{s=0}^{S_{\tilde{v}}} \alpha(s,r)$. A similar result is found in Lauritzen (1995). Recall that a local model without functional restrictions complies with a local part of a recursive graphical model.

For situations of functional restrictions in a local model, we typically have to carry out the maximization by an iterative method. Being able to calculate both first and second order derivatives for $Q^*(\theta_{\tilde{v}|r}|\theta^n, y)$, the globally convergent algorithm for maximization, as described in Jensen *et al.* (1991, Theorem 3), can be applied. The overall computational efficiency of the EM algorithm can be improved if only the first most influential iterations are considered here.

The acceleration algorithm

Recall that the generalized gradient for the posterior likelihood is approximated by an EM-step. Hence, to accomplish the specialization of the acceleration algorithm we only need to derive the gradient. It can be divided as specified in (3).

First consider the derivation of the gradient for the (traditional) log-likelihood of a single incomplete observation, denoted by y. From (1) we see that

$$\frac{\partial}{\partial \theta_{\tilde{v}|r}} \log p(y|\theta) = \frac{1}{p(y|\theta)} \sum_{x\in\mathcal{X}(y)} \frac{\partial}{\partial \theta_{\tilde{v}|r}} p(x|\theta). \qquad (9)$$

By global and local variation independence and by using the chain rule for differentiation

$$\begin{aligned} &\frac{\partial}{\partial \theta_{\tilde{v}|r}} p(x|\theta) \\ &\quad= \sum_{v\in\tilde{v}} \frac{p(x|\theta)}{p(x_v|x_{pa(v)}, \theta_{\tilde{v}|r})} \frac{\partial}{\partial \theta_{\tilde{v}|r}} p(x_v|x_{pa(v)}, \theta_{\tilde{v}|r}) \\ &\quad= p(x|\theta) \sum_{v\in\tilde{v}} \chi^r(x_{pa(v)}) \frac{\partial}{\partial \theta_{\tilde{v}|r}} \log p(x_v|r, \theta_{\tilde{v}|r}), \end{aligned}$$

where $\chi^r(x_{pa(v)})$ is defined similarly to (7).

Let $\tau(\theta_{\tilde{v}|r})$ denote the expected value of the statistic for the local exponential model defining $p(x_v|r, \theta_{\tilde{v}|r})$. By inserting the exponential representation we get

$$\frac{\partial}{\partial \theta_{\tilde{v}|r}} p(x|\theta) = p(x|\theta) \sum_{v\in\tilde{v}} \chi^r(x_{pa(v)}) \left(t(x_v) - \tau(\theta_{\tilde{v}|r}) \right). \qquad (10)$$

Finally, by inserting (10) into (9), the local components of the gradient for a single observation are derived as

$$\begin{aligned}\frac{\partial}{\partial\theta_{\bar{v}|r}} & \log p(y|\theta) \\ &= \sum_{x\in\mathcal{X}(y)} \frac{p(x|\theta)}{p(y|\theta)} \sum_{v\in\bar{v}} \chi^{r}(x_{pa(v)}) \left(t(x_v) - \tau(\theta_{\bar{v}|r})\right) \\ &= \sum_{v\in\bar{v}} \sum_{s=0}^{S_{\bar{v}}} p(s,r|y,\theta) \left(t(s) - \tau(\theta_{\bar{v}|r})\right).\end{aligned}$$

For a sample of independent observations the gradients for each observation simply add up. Hence, if y denotes a sample, then the gradient for the log-likelihood is given by the local components derived as

$$\frac{\partial}{\partial\theta_{\bar{v}|r}} l(\theta|y) = \sum_{s=0}^{S_{\bar{v}}} \tilde{n}(s,r) \left(t(s) - \tau(\theta_{\bar{v}|r})\right). \tag{11}$$

When functional restrictions are not specified in a local model, the gradient for the associated local log-prior is found by straightforward differentiation of the logarithm to the prior in (5), whereby

$$\frac{\partial}{\partial\theta_{\bar{v}|r}} l(\theta) = \sum_{s=0}^{S_{\bar{v}}} \alpha(s,r) \left(t(s) - \tau(\theta_{\bar{v}|r})\right). \tag{12}$$

Similarly, when restrictions are specified, the local gradient is derived by differentiation of the logarithm to the normal prior distribution in (6) as

$$\frac{\partial}{\partial\theta_{\bar{v}|r}} l(\theta) = -\beta\nu(\theta^{*}_{\bar{v}|r})(\theta_{\bar{v}|r} - \theta^{*}_{\bar{v}|r}). \tag{13}$$

Local computation of the posterior gradient, with components composed of (11) and one of (12) and (13), hereby implies that also the acceleration algorithm can be evaluated locally.

References

Andreassen, S., Jensen, F. V., Andersen, S. K., Falck, B., Kjærulff, U., Woldbye, M., Sørensen, A., Rosenfalck, A. & Jensen, F. (1989). MUNIN - an expert EMG assistant, *in* J. E. Desmedt (ed.), *Computer-Aided Electromyography and Expert Systems*, Elsevier Science Publishers, chapter 21, pp. 255–277.

Buntine, W. L. (1995). A guide to the literature on learning graphical models, *IEEE Transactions on Knowledge and Data Engineering* . Submitted.

Dempster, A. P., Laird, N. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society, Series B* **39**: 1–38.

Green, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation, *Journal of the Royal Statistical Society, Series B* **52**: 443–452.

Heckerman, D., Geiger, D. & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning* . To appear.

Jamshidian, M. & Jennrich, R. I. (1993). Conjugate gradient acceleration of the EM algorithm, *Journal of the American Statistical Association* **88**(421): 221–228.

Jensen, S. T., Johansen, S. & Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function, *Biometrika* **78**(4): 867–877.

Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data, *Computational Statistics & Data Analysis* **19**: 191–201.

Lauritzen, S. L. & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *Journal of the Royal Statistical Society, Series B* **50**: 157–224.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society, Series B* **44**(2): 226–233.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *Journal of the Royal Statistical Society, Series B* **51**(1): 127–138.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Series in Representation and Reasoning, Morgan Kaufmann Publishers, Inc.

Russell, S., Binder, J., Koller, D. & Kanazawa, K. (1995). Local learning in probabilistic networks with hidden variables, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* To appear.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. G. (1993). Bayesian analysis in expert systems, *Statistical Science* **8**(3): 219–247.

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

Thiesson, B. (1995). Score and information for recursive exponential models with incomplete data. Manuscript.

Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables, *Biometrika* **70**: 537–552.