Didelez, Pigeot:

Maximum Likelihood Estimation in Graphical Models with Missing Values

Sonderforschungsbereich 386, Paper 75 (1997)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Maximum Likelihood Estimation in Graphical Models with Missing Values

By Vanessa Didelez and Iris Pigeot

University of Munich, Institute of Statistics, Ludwigstr. 33, D-80539 Munich, Germany

Summary

In this paper we discuss maximum likelihood estimation when some observations are missing in mixed graphical interaction models assuming a conditional Gaussian distribution as introduced by Lauritzen & Wermuth (1989). For the saturated case ML estimation with missing values via the EM algorithm has been proposed by Little & Schluchter (1985). We expand their results to the special restrictions in graphical models and indicate a more efficient way to compute the E–step. The main purpose of the paper is to show that for certain missing patterns the computational effort can considerably be reduced.

Some key words: EM algorithm; Graphical interaction models; Maximum likelihood estimation; Missing pattern; Missing values.

1. Introduction

Graphical models are used to describe complex multivariate association structures. They are mainly of interest in empirical research in the social, psychological or behavioural sciences where a large number of variables is typically collected via questionnaire or interview. When analysing such data sets it is of interest to get to know associations between pairs of variables where usually it is not sufficient to allow for pure response and pure explanatory variables. In contrast, the association structure is such complex that so–called intermediates have to be introduced which are responses for some of the explanatory variables and explanatory for the responses and other intermediates. In such situations, more sophisticated models than simple regressions are called for. Graphical models have been developed to

cope with association structures of such a high complexity.
As mentioned above, the data are usually collected via questionnaire or interview. This gives rise to another problem. Here, missing values are very likely to occur because people refuse to answer or cannot remember the event which is asked for. Thus, it is an essential task to find procedures which are on the one hand adequate for estimating the parameters of a graphical model in presence of missing values and on the other hand easy to handle. We focus here on maximum likelihood (ML) estimation in mixed graphical interaction models assuming a conditional Gaussian (CG) distribution where ML estimation typically requires iterative solutions and thus appropriate algorithms. Missing patterns which allow for simplifications and efficient computation are therefore of special concern.
The outline of the paper is as follows. In Section 2 we give a short introduction to graphical interaction models with CG distribution. Some of the most important properties of such models are reviewed. ML estimates are presented for the saturated model and in the special case of a G–Markovian CG distribution. The application of the EM algorithm for calculating the ML estimates in case that the missing values occur at random is discussed in Section 3. Since computational effort can be quite high, Section 4 emphasizes possibilities for simplifying the algorithm dealing with special missing patterns which make computation much easier when being taken into account. An example is given for illustrating the gained reduction in computational effort. Additional aspects are addressed in the discussion.

2. Graphical Models and ML Estimation with complete data

For convenience let us briefly introduce graphical interaction models with CG distribution using the terminology established by Lauritzen & Wermuth (1989). Consider a random vector $X = (Y^\top, I^\top)^\top$ where $Y = (Y_1, \ldots, Y_R)^\top$ is a vector of R continuous variables with realizations $y \in I\!R^R$ and $I = (I_1, \ldots, I_Q)^\top$ a vector of Q discrete variables with $\mathcal{I}$ denoting the set of possible realizations i. The vector X is said to have a CG distribution if the density function $f(x)$ is given by

$$f(x) = f(y, i) = p(i)\varphi(y|\mu(i), \Sigma(i)),$$

where $p(i)$ is the discrete marginal probability of $I = i$ with $p(i) > 0$ for all $i \in \mathcal{I}$ and $\varphi(\cdot|\mu(i), \Sigma(i))$ is the density of a multivariate normal distribution with mean

vector $\mu(i) \in I\!R^R$ and covariance matrix $\Sigma(i) \in I\!R^{R\times R}$. We assume $\Sigma(i)$ to be positive definite for all $i \in \mathcal{I}$. The set $\{p(i), \mu(i), \Sigma(i) | i \in \mathcal{I}\}$ represents the moment parameterisation of the CG distribution and can be transformed to the canonical parameters $\{g(i), h(i), K(i) | i \in \mathcal{I}\}$ by

$$\begin{aligned} g(i) &= \log p(i) - \frac{|\Gamma|}{2}\log(2\pi) - \frac{1}{2}\left(\log|\Sigma(i)| + \mu(i)^\top \Sigma(i)\mu(i)\right), \\ h(i) &= \Sigma(i)^{-1}\mu(i) \quad \text{and} \quad K(i) = \Sigma(i)^{-1}. \end{aligned}$$

The set $\{p(i), h(i), K(i) | i \in \mathcal{I}\}$ is called the standard mixed characteristics and is often most convenient. The graphical models we would like to consider specify conditional independencies between certain components of the vector X which can be represented by a graph and which result in restrictions on the parameters, usually formulated for the canonical parameters (see Lauritzen & Wermuth, 1989). A graph $G = (V, E)$ is given by a nonempty finite set V of vertices and a set $E \subseteq V \times V$ of edges. We only consider undirected graphs, that is $(a, b) \in E \Rightarrow (b, a) \in E$. If we identify the set of indices of the components of X with V then a multivariate distribution is called G–Markovian if it holds the following conditional independencies

$$X_a \perp X_b | X_{V\setminus\{a,b\}} \quad \text{for all } (a, b) \notin E,\ a \neq b. \tag{1}$$

Property (1) is called the pairwise Markov property. For CG distributions (1) is equivalent to the global Markov property, that is for $A, B, D \subset V : X_A \perp X_B | X_D$ whenever D separates A and B in G (Lauritzen & Wermuth, 1989) where $X_A = (X_a)_{a\in A}$ for $A \subseteq V$. If $E = V \times V$ the graph is called complete and the corresponding graphical model is the saturated one since there are no conditional independencies. If G is not complete maximal subsets of V without any pairwise conditional independencies are called cliques, that is $C \subseteq V$ is a clique of G if (a) for all $a, b \in C, a \neq b$ it follows that $(a, b) \in E$ and (b) for all $a \in V\setminus C$ it follows that there exists $b \in C$ with $(a, b) \notin E$. The cliques of a graph are unique.
For $A \subset V$ the induced subgraph is defined by $G_A = (A, E_A)$ with $E_A = E \cap (A \times A)$. We will further denote by $\mathcal{M}(G)$ the statistical model containing all G–Markovian CG distributions. It will be necessary to distinguish between $\mathcal{M}(G)_A$ which denotes the set of A–marginals of all G–Markovian CG distributions and $\mathcal{M}(G_A)$ which denotes the set of all G_A–Markovian CG distributions. In general they are not the same. Let in addition $\mathcal{M}(G)^A$ be the set of conditional G–Markovian CG distributions conditioning on the variables X_A.

As X contains continuous and discrete components it will be convenient to divide the index set V into disjoint sets $V = \Gamma \dot{\cup} \Delta$ where Δ is the index set of the discrete components and Γ that of the continuous ones. Note that Δ or Γ may be empty yielding a graphical model with multivariate normal distribution or a loglinear graphical model, respectively. Whenever Γ and Δ are nonempty the corresponding graph is called a marked graph.

In order to decompose the usually rather complex estimation problem in graphical models into "smaller" estimation problems we will further need the notions of decomposability and collapsibility. A graph G is collapsible onto a subset $A \subset V$ if $B = V \backslash A$ is a strong and simplicial collection that is if each connected component $B_k, k = 1, \ldots, K$, of B holds (a) $\mathrm{bd}(B_k)$ is complete and (b) $B_k \subseteq \Gamma$ or $\mathrm{bd}(B_k) \subseteq \Delta$ where $\mathrm{bd}(B_k)$ is the boundary of B_k that is the set of vertices $b \in V \backslash B_k$ with $(a, b) \in E$ for one $a \in B_k$. Given that the joint distribution of X is from the class of G–Markovian distributions collapsibility is equivalent to the class of marginal distribution of X_A being identical to the class of G_A–Markovian distributions (Frydenberg, 1990). A decomposition $A \dot{\cup} B \dot{\cup} D = V$ of a marked graph G is defined by (a) D separates A and B, (b) D is complete and (c) $D \subseteq \Delta$ or $B \subseteq \Gamma$. Given such a decomposition (A, B, D) the graph is collapsible onto $A \cup D$. A graph is decomposable if it is complete or if there exists a decomposition (A, B, D) with A and B both nonempty into decomposable subgraphs $G_{A \cup D}$ and $G_{B \cup D}$. Decomposability can be checked by verifying that the graph is triangulated and does not contain any path between two discrete vertices passing through only continuous vertices with the discrete vertices not being neighbours. For decomposable graphs there always exist closed expressions for the ML estimates of the parameters of the corresponding CG distribution (Leimer, 1989, Frydenberg & Lauritzen, 1989).

Given a random sample $X^1, \ldots, X^N$ of i.i.d. random vectors where $X^j = ({Y^j}^\top, {I^j}^\top)^\top$ is distributed according to a CG distribution the set of joint distributions constitutes an exponential family with sufficient statistics $N(i) = \sum_{j=1}^{N} \chi(I^j = i)$, $S(i) = \sum_{j \in \mathcal{J}(i)} Y^j$ and $SS(i) = \sum_{j \in \mathcal{J}(i)} Y^j {Y^j}^\top$ for $i \in \mathcal{I}$, where χ is the indicator function and $\mathcal{J}(i) = \{j \in \{1, \ldots, N\} | i^j = i\}$. The realized sufficient statistics are denoted by $n(i), s(i)$ and $ss(i)$. For complete data the ML estimates in the saturated model are given by

$$\hat{p}(i) = \frac{n(i)}{N}, \qquad \hat{\mu}(i) = \bar{y}(i) = \frac{s(i)}{n(i)} \quad \text{and} \quad \hat{\Sigma}(i) = ssd(i) = \frac{ss(i)}{n(i)} - \hat{\mu}(i)\hat{\mu}(i)^\top,$$

for $i \in \mathcal{I}$. They exist with probability one when $n(i) > R$ for all $i \in \mathcal{I}$. If in addition the CG distribution is G–Markovian with respect to a graph G that is not complete, the set of sufficient statistics reduces as follows (Lauritzen, 1996). Let $\mathcal{C}_\Delta$ denote the set of cliques in the graph induced by the discrete vertices; let further $\mathcal{C}_\Delta(r), r \in \Gamma$, be the sets $d \subseteq \Delta$ with $d \cup \{r\}$ a clique in $G_{\Delta \cup \{r\}}$ and $\mathcal{C}_\Delta(r,s), r,s \in \Gamma$, the sets $d \subseteq \Delta$ with $d \cup \{r,s\}$ a clique in $G_{\Delta \cup \{r,s\}}$. Then the minimal sufficient statistics are

(i) the marginal tables of counts $N(i_d) = \sum_{j=1}^N \chi(I_d^j = i_d)$, $d \in \mathcal{C}_\Delta$,

(ii) for each continuous variable $r \in \Gamma$ the set of marginal tables of sums $S(i_d)_r = \sum_{j \in \mathcal{J}(i_d)} Y_r^j$ and sums of squares $SS(i_d)_r = \sum_{j \in \mathcal{J}(i_d)} (Y_r^j)^2$, $d \in \mathcal{C}_\Delta(r)$,

(iii) for each edge (r,s), $r \neq s$, between continuous variables the marginal tables of sums of products $SS(i_d)_{r,s} = \sum_{j \in \mathcal{J}(i_d)} Y_r^j Y_s^j$, $d \in \mathcal{C}_\Delta(r,s)$.

Note that the sufficient statistics of the restricted model are sums of the sufficient statistics of the saturated model. The ML estimates are given by equating these sufficient statistics with their expectations. No general conditions to guarantee the existence of the ML estimates can be given except for special cases for example when there exists a decomposition or when the graph is decomposable which we consider next.

Given a decomposition (A,B,D) of G the graph is collapsible onto $A \cup D$ and the joint density factorizes as follows

$$f(x) = f(x_{A \cup D}) f(x_B | x_D), \tag{2}$$

where $f(x_{A \cup D})$ and $f(x_B|x_D)$ denote the marginal and conditional densities of $X_{A \cup D}$ and $X_B|X_D$, respectively. Let θ denote the parameter vector of the joint density, that is $\theta = (p(i), \mu(i), \Sigma(i))_{i \in \mathcal{I}}$, and let $\theta_{A \cup D}, \theta_{B|D}$ denote the corresponding parameter vectors of the marginal and conditional densities which both are densities of CG distributions. Using (2) the log–likelihood $L(\theta|x)$ is

$$L(\theta|x) = \sum_{j=1}^N \log f(x_{A \cup D}^j | \theta_{A \cup D}) + \sum_{j=1}^N \log f(x_B^j | x_D^j; \theta_{B|D}),$$

where $L(\theta|x)$ can be maximized by separately maximizing the two summands. It follows from the central result of Frydenberg & Lauritzen (1989, Proposition 4) that the first is maximized by the ML estimate in $\mathcal{M}(G_{A \cup D})$ based upon data $(x_{A \cup D}^1, \ldots, x_{A \cup D}^N)$ and the second by the ML estimate in the regression model

$\mathcal{M}(G_{B\cup D})^D$ based upon data $(x^1_{B\cup D},\ldots,x^N_{B\cup D})$. The estimation in $\mathcal{M}(G_{B\cup D})^D$ in turn is based on the estimates in $\mathcal{M}(G_{B\cup D})$ and $\mathcal{M}(G_D)$. Let $(\hat{p}_{[C]},\hat{h}_{[C]},\hat{K}_{[C]})$ denote the ML estimates of the standard mixed characteristics in the model $\mathcal{M}(G_C)$ for any $C\subseteq V$, and let $\{M\}^0$ be the matrix or vector obtained from M by filling up with zero entries so as to give it full dimension $R\times R$ or R. If (A,B,D) is a decomposition of G it is shown by the same authors that the ML estimates of the standard mixed characteristics in $\mathcal{M}(G)$ are given by

$$\hat{p}(i) = \frac{\hat{p}_{[A\cup D]}(i_{A\cup D})\hat{p}_{[B\cup D]}(i_{B\cup D})}{\hat{p}_{[D]}(i_D)}, \tag{3}$$

$$\hat{h}(i) = \{\hat{h}_{[A\cup D]}(i_{A\cup D})\}^0+\{\hat{h}_{[B\cup D]}(i_{B\cup D})\}^0-\{\hat{h}_{[D]}(i_D)\}^0, \tag{4}$$

$$\hat{K}(i) = \{\hat{K}_{[A\cup D]}(i_{A\cup D})\}^0+\{\hat{K}_{[B\cup D]}(i_{B\cup D})\}^0-\{\hat{K}_{[D]}(i_D)\}^0, \tag{5}$$

where for any $C\subseteq V$: $i_C = i_{C\cap\Delta}$. These results can be applied to the situation of a graph G being collapsible onto a set A (Frydenberg, 1990) using that $(V\backslash\mathrm{cl}(B_k),B_k,\mathrm{bd}(B_k))$ is a decomposition of G for every $k=1,\ldots,K$, where $B=V\backslash A$ and $B_1,\ldots,B_K$ are the connected components of B and that the joint density factorizes as

$$f(x)=f(x_A)f(x_{B_1}|x_{\mathrm{bd}(B_1)})\cdots f(x_{B_K}|x_{\mathrm{bd}(B_K)}).$$

In addition Frydenberg & Lauritzen (1989) show that closed expressions of the ML estimates exist for decomposable graphs. In general iterative procedures are needed to calculate the ML estimates (Frydenberg & Edwards, 1989).

3. Application of the EM algorithm

As already mentioned it often occurs that a collected data set is incomplete. This means that for some sample entities some of the components of the observation vector are missing. Thus, we can divide each observation vector into its observed and missing components, i.e. $X=(X^\top_{\mathrm{Obs}},X^\top_{\mathrm{Mis}})^\top=(Y^\top_{\mathrm{Obs}},I^\top_{\mathrm{Obs}},Y^\top_{\mathrm{Mis}},I^\top_{\mathrm{Mis}})^\top$. Note that the sets of observed and missing variables can be different for each observation vector X^j, $j=1,\ldots,N$. To reduce computational effort it will be helpful to process all cases with identical missing pattern in the same step if such cases exist. In the following we assume that for every entity at least one component of X can be observed.

In addition we assume missingness at random (MAR) i.e. the missing mechanism is conditionally independent of the missing value given the observed components, it may depend on the latter ones. This strong assumption should be carefully verified in practice since violations of the MAR assumption can lead to considerable bias of the estimates. Under MAR, however, it is possible to get the ML estimates without any further knowledge about the missing mechanism (Rubin, 1974). Their calculation requires maximization of the likelihood of the observed variables. This can be a tedious task especially in complex multivariate models as the ones considered here where even with complete data the ML estimates do not always exist in closed form. A general tool for handling this problem is the EM algorithm (Dempster, Laird & Rubin, 1977) which is easy to apply when the considered model is an exponential family. This algorithm consists of two steps: the E–step that calculates the expected sufficient statistics given the observed data and the current estimates of the parameters, and the M–step that determines the new estimates using the conditional expectations of the sufficient statistics as if they were the observed. Its drawback is its slow convergence rate wherefore alternative strategies are worthwhile to explore.

We start by describing the EM algorithm for mixed interaction models with CG distribution. The conditional expectations of the sufficient statistics given the observed values are as follows

$$\text{(i)} \qquad E(N(i_d)|x_{\text{Obs}}) = \sum_{j=1}^{N} \text{pr}(I_d = i_d|x^j_{\text{Obs}})$$

$$\text{(ii)} \qquad E(S(i_d)_r|x_{\text{Obs}}) = \sum_{j=1}^{N} \text{pr}(I_d = i_d|x^j_{\text{Obs}})E(Y_r|y^j_{\text{Obs}}, i_d) \quad \text{and}$$

$$E(SS(i_d)_r|x_{\text{Obs}}) = \sum_{j=1}^{N} \text{pr}(I_d = i_d|x^j_{\text{Obs}})[(E(Y_r|y^j_{\text{Obs}}, i_d))^2 + \text{var}(Y_r|y^j_{\text{Obs}}, i_d)],$$

$$\text{(iii)} \qquad E(SS(i_d)_{r,s}|x_{\text{Obs}})$$
$$= \sum_{j=1}^{N} \text{pr}(I_d = i_d|x^j_{\text{Obs}})[E(Y_r|y^j_{\text{Obs}}, i_d)E(Y_s|y^j_{\text{Obs}}, i_d) + \text{cov}(Y_r, Y_s|y^j_{\text{Obs}}, i_d)].$$

A first approach to calculate these conditional expectations will be to extend the results of Little & Schluchter (1985) as already indicated by Edwards (1996). This means that the conditional expectations of the sufficient statistics of the saturated model have to be computed and that those of the restricted model are then obtained

by appropriate summation over the former ones. Note that

$$\mathrm{pr}(I_d = i_d | x_{\mathrm{Obs}}) = \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu(i'),$$

where $\nu(i) = \mathrm{pr}(I = i | x_{\mathrm{Obs}})$ is the posterior probability for an observation to lie in cell i given all its observed components. To compute the posteriors let $(\mu(i)_{\mathrm{Obs}}, \Sigma(i)_{\mathrm{Obs}})$ be the parameters of the marginal distribution of Y_{Obs} given $I = i$. Let further denote $\mathcal{S} = \{(i_{\mathrm{Obs}}, i_{\mathrm{Mis}}) | i_{\mathrm{Mis}} \in \mathcal{I}_{\mathrm{Mis}}\}$ the set of cells the observation could lie in given the observed discrete components. Then

$$\nu(i) = \frac{\exp \kappa(i)}{\sum_{s \in \mathcal{S}} \exp \kappa(s)}$$

with

$$\begin{aligned} \kappa(i) = & \, y_{\mathrm{Obs}}^\top \Sigma(i)_{\mathrm{Obs}}^{-1} \mu(i)_{\mathrm{Obs}} \\ & -\frac{1}{2} \left[y_{\mathrm{Obs}}^\top \Sigma(i)_{\mathrm{Obs}}^{-1} y_{\mathrm{Obs}} + \mu(i)_{\mathrm{Obs}}^\top \Sigma(i)_{\mathrm{Obs}}^{-1} \mu(i)_{\mathrm{Obs}} \right] + \log p(i). \end{aligned}$$

This slightly differs from the formulae given by Little & Schluchter (1985) since due to the non–homogeneity assumption the term $\frac{1}{2} y_{\mathrm{Obs}}^\top \Sigma(i)_{\mathrm{Obs}}^{-1} y_{\mathrm{Obs}}$ does not cancel out. Note that $\nu(i) = 0$ if $i \notin \mathcal{S}$ and $\nu(i) = 1$ if $\mathcal{S} = \{i\}$.

In addition, we need the conditional expectation $E(Y_r | y_{\mathrm{Obs}}, i)$ and the conditional covariance $\mathrm{cov}(Y_r, Y_s | y_{\mathrm{Obs}}, i)$ for missing continuous components Y_r, Y_s. They can easily be computed for given parameters $\mu(i)$ and $\Sigma(i)$ using the properties of the multivariate normal distribution:

$$\begin{aligned} E(Y_r | y_{\mathrm{Obs}}, i) &= \mu(i)_r - \Sigma(i)_{\{r\},\mathrm{Obs}} \Sigma(i)_{\mathrm{Obs}}^{-1} (y_{\mathrm{Obs}} - \mu(i)_{\mathrm{Obs}}) = y_r(i), \\ \mathrm{cov}(Y_{\{r,s\}} | y_{\mathrm{Obs}}, i) &= \Sigma(i)_{\{r,s\}} - \Sigma(i)_{\{r,s\},\mathrm{Obs}} \Sigma(i)_{\mathrm{Obs}}^{-1} \Sigma(i)_{\mathrm{Obs},\{r,s\}}, \end{aligned}$$

where $\mathrm{cov}(Y_{\{r,s\}} | y_{\mathrm{Obs}}, i)$ denotes the conditional covariance matrix of $Y_{\{r,s\}}$ with entries $\mathrm{cov}(Y_r, Y_s | y_{\mathrm{Obs}}, i)$ as conditional covariance of Y_r and Y_s, $\mathrm{var}(Y_r | y_{\mathrm{Obs}}, i)$ and $\mathrm{var}(Y_s | y_{\mathrm{Obs}}, i)$ each as conditional variance. These entries will be denoted by $c_{r,s}(i)$. If the continuous components are not missing, we get $y_r(i) = y_r$ and $c_{r,s}(i) = 0$.

Now we can compute the conditional expectations of the sufficient statistics given the observed data in a graphical model with CG distribution following a graph G. They are given as follows

$$E(N(i_d) | x_{\mathrm{Obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i'), \qquad d \in \mathcal{C}_\Delta, \tag{6}$$

where $\nu^j(i)$ is $\nu(i)$ for the j–th observation,

$$E(S(i_d)_r|x_{\text{Obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i') y_r^j(i'), \qquad d \in \mathcal{C}_\Delta(r),\ r \in \Gamma, \tag{7}$$

and for $r = s$ or $(r, s) \in E$ $(r, s \in \Gamma)$:

$$E(SS(i_d)_{r,s}|x_{\text{Obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i')[y_r^j(i') y_s^j(i') + c_{r,s}^j(i')], \qquad d \in \mathcal{C}_\Delta(r, s). \tag{8}$$

The E–step of the EM algorithm determines (6), (7) and (8) for the current parameter iterates. While (6) and (7) differ from the saturated case only through the additional summation over $i' \in \mathcal{I} : i'_d = i_d$ and thus constitute no simplification we can see from (8) that the conditional covariances only have to be calculated for missing continuous components Y_r and Y_s with $(r, s) \in E$. The M–step consists in computing the next parameter iterates using the conditional expectations of the sufficient statistics as if they were the observed ones and thus can be performed in the same way as for complete data.

As noticed by Lauritzen (1995) the effort to compute the E–step can be considerable in particular when dealing with high dimensions. Considering the following example it becomes clear that an acceleration is possible. To compute the conditional expectation of the sufficient statistics $E(N(i_d)|X_{\text{Obs}})$ we need the probabilities $\text{pr}(I_d = i_d|x_{\text{Obs}})$. If now the set of observed variables x_{Obs} contains the boundary of d, then it follows from the local Markov property that $\text{pr}(I_d = i_d|x_{\text{Obs}}) = \text{pr}(I_d = i_d|x_{\text{bd}(d)})$ what makes clear that the computation depends on fewer variables. If in contrast there is no path in G from the set d to the set of observed variables then we even have marginal independence and $\text{pr}(I_d = i_d|x_{\text{Obs}}) = \text{pr}(I_d = i_d)$. These are cases in which the computation can be simplified but that are not taken into account if we proceed as described above. The procedure proposed by Lauritzen (1995) to accelerate the E–step relies on a computational scheme developed by Lauritzen & Spiegelhalter (1988) in the context of probabilistic expert systems. Lauritzen (1995) considers graphical models with only discrete variables but points out that the procedure can be generalized to work for mixed graphical interaction models using the propagation scheme of Lauritzen (1992). The mentioned probabilistic expert systems specify the existing knowledge about association structures in a system of variables by graphical models. For given evidence, that is for known values of a subset of the variables,

properties of the updated system are of interest where updating corresponds to a conditioning process. The computational task is therefore essentially the same as for the E–step if we consider the observed values as evidence and the conditional expectations of the sufficient statistics as interesting properties. The possible gain in computational ease is based on two aspects. Computation can be done with unnormalized density functions and the Markov properties of the graph can be exploited being reflected by the product structure of the joint density. For this it is necessary to form a junction tree that is a special organization of the cliques of the graph so that calculations can rely on operations only between neighbouring cliques. The operations in turn are done on CG potentials avoiding normalization. For further details we refer to Lauritzen (1992).

4. Special Missing Patterns

The EM–algorithm applies when the marginal likelihood of the observed data is too complicated to be maximized directly. In some situations, however, we can find simple formulae for this marginal likelihood by factorization. This is well known for monotone missing patterns and certain underlying distributions as the multinomial and multivariate Gaussian (Little & Rubin, 1987). These distributions have the property that their conditional and marginal distributions are of the same type. The joint likelihood can be factorized by suitable conditional and marginal densities allowing a separate maximization of each factor. Given a monotone missing pattern, we can then find a factorization in these models such that maximization of each factor corresponds to a complete data situation. In general this simplification only works for saturated models because maximizing separately is often impossible when there are restrictions on the parameters. A lot of papers in the literature on graphical models, however, are concerned with simplifications of the estimation problem using the properties of decomposability and collapsibility of graphs leading to factorizations of the likelihood. As shown in Section 2 a decomposition of a graph leads to ML estimates that are functions of the ML estimates in special submodels induced by the decomposing sets. We will now make use of factorization (2) to show that for a special missing pattern the computation of the ML estimates needs no further effort than for complete data. In addition we will indicate more general

missing patterns for which at least a separate application of the EM algorithm to the submodels generated by $A \cup D$ and $B \cup D$ is possible yielding the ML estimates in $\mathcal{M}(G)$ in a similar way as given by (3), (4) and (5).

Let (A, B, D) be a decomposition of the graph G. The missing pattern that will be of interest here is given when only the components X_B are incompletely observed in a way that the whole vector X_B is either missing or observed. Let $V^B \subseteq \{1, \ldots, N\}$, $V^B \neq \emptyset$, denote the index set of those observations where X_B is known. It follows that for an incomplete observation the marginal density of the observed variables is $f(x_{A\cup D})$ and the log–likelihood of the observed data is thus given by

$$L_{\text{Obs}}(\theta|x_{\text{Obs}}) = \sum_{j=1}^{N} \log f(x^j_{A\cup D}|\theta_{A\cup D}) + \sum_{j\in V^B} \log f(x^j_B|x^j_D;\theta_{B|D}). \tag{9}$$

It follows that we can get the ML estimates of $\theta_{A\cup D}$ and $\theta_{B|D}$ by separately maximizing

$$\begin{aligned} L_{A\cup D}(\theta_{A\cup D}|x^j_{A\cup D}; j = 1, \ldots, N) &= \sum_{j=1}^{N} \log f(x^j_{A\cup D}|\theta_{A\cup D}), \\ L_{B|D}(\theta_{B|D}|x^j_{B\cup D}; j \in V^B) &= \sum_{j\in V^B} \log f(x^j_B|x^j_D;\theta_{B|D}), \end{aligned}$$

where both log–likelihoods correspond to complete data situation, the maximization of $L_{B|D}$ being based on a smaller data set with the observations $j \in V^B$. The ML estimates for this incomplete data situation are given as in (3),(4) and (5) replacing $\hat{p}_{[B\cup D]}$, $\hat{p}_{[D]}$, $\hat{h}_{[B\cup D]}$, $\hat{h}_{[D]}$, $\hat{K}_{[B\cup D]}$ and $\hat{K}_{[D]}$ by the estimates based only on the observations $j \in V^B$.

If we consider more general missing patterns it will be necessary to fall back upon the EM algorithm. But a special missing pattern will at least allow a separate application of the algorithm in the models $\mathcal{M}(G_{A\cup D})$ and $\mathcal{M}(G_{B\cup D})$. This pattern is given whenever X_D is "more observed" than X_A and X_B. To describe this formally let Obs(A) denote the observed components of a subvector X_A for any $A \subseteq V$. In a sample $X^1, \ldots, X^N$ the vector X_D is more observed than X_A if from Obs$(A) \neq \emptyset$ it follows that Obs$(D) = D$ for each observation. To see why this pattern allows a separate application of the EM algorithm let us recall that for a decomposition (A, B, D) the conditional independence $A \perp B|D$ holds. Thus, if components of X_A are missing and X_D is fully observed the conditional expectation of a function depending on $X_{A\cup D}$ given the observed variables X_{Obs} is identical with the conditional

expectation of the same function given $(X_{\mathrm{Obs}(A)}, X_D)$, where knowledge of $X_{\mathrm{Obs}(B)}$ is not required. The same argument holds for functions of $X_{B\cup D}$ if components of X_B are missing and again X_D is completely observed. This can directly be applied in the E–step of the EM algorithm since the sufficient statistics are either functions of $X_{A\cup D}$ or $X_{B\cup D}$ if D separates A and B in G. This yields that both, E–step and M–step, can be processed separately. The observations where X_A and X_B are completely missing only contribute to the estimation of the parameters in $\mathcal{M}(G_{A\cup D})$ since they contain no information about the conditional distribution of X_B given X_D. More formally, let $V^A, V^B \subseteq \{1,\ldots,N\}$, both nonempty, denote the index sets of those observations for which at least one component of X_A and X_B, respectively, is known, and $V^{\overline{AB}}$ those for which all components $X_{A\cup B}$ are missing but at least one of X_D is observed. The set $V^{\overline{AB}}$ may be empty and it is possible that $V^A = V^B$ or $V^A \cap V^B = \emptyset$. The factorization (2) yields similarly as in (9) the log–likelihood of the observed data

$$L_{\mathrm{Obs}}(\theta|x_{\mathrm{Obs}}) = \sum_{j\in V^A\cup V^{\overline{AB}}} \log f(x^j_{\mathrm{Obs}(A\cup D)}|\theta_{A\cup D}) + \sum_{j\in V^B} \log f(x^j_{\mathrm{Obs}(B)}|x^j_D;\theta_{B|D}),$$

where $f(x^j_{\mathrm{Obs}(A\cup D)}|\theta_{A\cup D})$ is the marginal density of the observed variables $X_{\mathrm{Obs}(A\cup D)}$ and $f(x^j_{\mathrm{Obs}(B)}|x^j_D;\theta_{B|D})$ the one of the variables $X_{\mathrm{Obs}(B)}$ given X_D. We then have the following result. The ML estimates of the mixed characteristics in model $\mathcal{M}(G)$ are given as in (3), (4) and (5) where

(i) $\hat{p}_{[A\cup D]}$, $\hat{h}_{[A\cup D]}$ and $\hat{K}_{[A\cup D]}$ result from maximization of the likelihood in model $\mathcal{M}(G_{A\cup D})$ based on data $(x^j_{\mathrm{Obs}(A)\cup D};\, j\in V^A)$ and $(x^j_{\mathrm{Obs}(D)};\, j\in V^{\overline{AB}})$,

(ii) $\hat{p}_{[B\cup D]}$, $\hat{h}_{[B\cup D]}$ and $\hat{K}_{[B\cup D]}$ result from maximization of the likelihood in model $\mathcal{M}(G_{B\cup D})$ based on data $(x^j_{\mathrm{Obs}(B)\cup D};\, j\in V^B)$ and

(iii) $\hat{p}_{[D]}$, $\hat{h}_{[D]}$ and $\hat{K}_{[D]}$ result from maximization of the likelihood in model $\mathcal{M}(G_D)$ based on data $(x_D;\, j\in V^B)$. Note that this corresponds to a complete data problem since X_D is always completely observed for $j\in V^B$.

Depending on the missing patterns within the vectors $X_{A\cup D}$, $X_{B\cup D}$ the separate maximizations in (i) and (ii) possibly require the EM algorithm. If we have a symmetric decomposition which means that the sets A and B are interchangeable as it is the case for pure graphs or when D contains all discrete variables the graph is collapsible onto D. For a missing pattern where D is more observed than A and B we can then

get the estimates by separate maximization in the models $\mathcal{M}(G_{A\cup D})^D$ based on data $j \in V^A$, $\mathcal{M}(G_{B\cup D})^D$ based on data $j \in V^B$ and $\mathcal{M}(G_D)$ based on all observations.

Concerning conditions related to the existence of the ML estimates we refer to Frydenberg & Lauritzen (1989). Of course, even when ML estimates exist for complete data this is not necessarily the case with missing values since problems of identification can occur. This has to be taken into account by choosing a sparser model if necessary.

5. Example

Following Frydenberg & Lauritzen (1989) let us consider the following graph for illustration: $G = (V, E)$ with $V = \{I_1, I_2, Y_1, Y_2\}$ and $E = V \times V \backslash \{(I_1, Y_2), (Y_2, I_1)\}$. The graphical representation is given in Fig. 1.

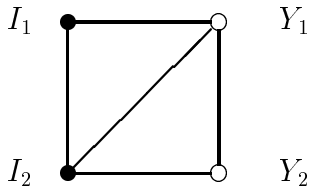

Fig. 1: A decomposable marked graph.

With $A = \{I_1\}$, $B = \{Y_2\}$ and $D = \{I_2, Y_1\}$ we have a decomposition and since A and B are complete the graph is decomposable. To determine the sufficient statistics note that $\mathcal{C}_\Delta = \{\{I_1, I_2\}\}$, $\mathcal{C}_\Delta(1) = \{\{I_1, I_2\}\}$, $\mathcal{C}_\Delta(2) = \{\{I_2\}\}$ and $\mathcal{C}_\Delta(1,2) = \{\{I_2\}\}$. Thus, the sufficient statistics are given by $N(i_1, i_2)$, $S(i_1, i_2)_1$, $SS(i_1, i_2)_1$, $S(i_2)_2$, $SS(i_2)_2$ and $SS(i_2)_{1,2}$ for $(i_1, i_2) \in \mathcal{I}$. Having complete data there exist explicit ML estimates that are given in Frydenberg & Lauritzen (1989) for the chosen graph. Let us now assume that Y_2 is incompletely observed. The application of the EM algorithm would require the computation of

$$E(Y_2|i^j, y_1^j) = \mu(i^j)_2 + \frac{\sigma(i^j)_{12}}{\sigma(i^j)_1}(y_1^j - \mu(i^j)_1) \quad \text{and}$$

$$\text{var}(Y_2|i^j, y_1^j) = \sigma(i^j)_2 - \frac{\sigma(i^j)_{12}}{\sigma(i^j)_1}$$

for each incomplete observation $j \in V \backslash V^B$ and for the current parameter iterates. Since the discrete variables I_1 and I_2 are completely observed we have either $\nu^j(i) = 1$ or $\nu^j(i) = 0$ depending on whether $I^j = i$ or not. These quantities are used to compute (6), (7) and (8) yielding the E–step. In the M–step the new parameter iterates are determined in the same way as for complete data. The EM algorithm is not complicated for this missing situation but taking into account that we have the special missing pattern that allows explicit estimates according to Section 4 avoids iterating. These explicit ML estimates are given as follows. Compute the ML estimates in the submodels $\mathcal{M}(G_{I_1,I_2,Y_1})$ using all observations and $\mathcal{M}(G_{I_2,Y_1,Y_2})$ as well as $\mathcal{M}(G_{I_2,Y_1})$ using only the complete observations and combine them according to (3), (4) and (5). In this special case we have

$$\begin{aligned}
\hat{p}_{[I_1,I_2,Y_1]}(i_1,i_2) &= \frac{n(i_1,i_2)}{N}, \\
\hat{K}_{[I_1,I_2,Y_1]}(i_1,i_2) &= n(i_1,i_2)[ssd_{[Y_1]}(i_1,i_2)]^{-1}, \\
\hat{h}_{[I_1,I_2,Y_1]}(i_1,i_2) &= \hat{K}_{[I_1,I_2,Y_1]}(i_1,i_2)\bar{y}_1(i_1,i_2)
\end{aligned}$$

estimated from all observations since the variables I_1, I_2 and Y_1 are always observed. The other estimates indexed by $[I_2, Y_1, Y_2]$ and $[I_2, Y_1,]$ make use only of the complete observations that is those for which Y_2 is observed. They will be denoted by * to mark the difference. We then have

$$\begin{aligned}
\hat{p}^*_{[I_2,Y_1,Y_2]}(i_2) &= \frac{n^*(i_2)}{N^*} \quad \text{and} \\
\hat{p}^*_{[I_2,Y_1]}(i_2) &= \frac{n^*(i_2)}{N^*},
\end{aligned}$$

where $n^*(i_2) = |\{j \in V^B | i_2^j = i_2\}|$ and $N^* = |V^B|$. And further

$$\begin{aligned}
\hat{K}^*_{[I_2,Y_1,Y_2]}(i_2) &= n^*(i_2)[ssd^*_{[Y_1,Y_2]}(i_2)]^{-1}, \\
\hat{h}^*_{[I_2,Y_1,Y_2]}(i_2) &= \hat{K}^*_{[I_2,Y_1,Y_2]}(i_2)\bar{y}^*(i_2), \\
\hat{K}^*_{[I_2,Y_1]}(i_2) &= n^*(i_2)[ssd^*_{[Y_1]}(i_2)]^{-1}, \\
\hat{h}^*_{[I_2,Y_1]}(i_2) &= \hat{K}^*_{[I_2,Y_1]}(i_2)\bar{y}^*_1(i_2).
\end{aligned}$$

Let us now consider a missing pattern where only I_2 and Y_1 are always observed that is I_1 and Y_2 are sometimes missing but not necessarily simultaneously. We then have the second type of missing pattern mentioned in Section 4 where the separating

set is more observed than the separated ones. The estimates indexed by $[I_2, Y_1, Y_2]$ and $[I_2, Y_1,]$ remain the same as above based on those observations where I_2, Y_1 and Y_2 are observed. They are not affected by the incompleteness of I_1. Of course, $\hat{p}_{[I_1,I_2,Y_1]}$, $\hat{h}_{[I_1,I_2,Y_1]}$ and $\hat{K}_{[I_1,I_2,Y_1]}$ are affected. Here, the data base cannot be reduced to those observations where I_1, I_2 and Y_1 are completely known since then the information from the observations where only I_2 and Y_1 are observed would be lost. The estimation of $p_{[I_1,I_2,Y_1]}$, $h_{[I_1,I_2,Y_1]}$ and $K_{[I_1,I_2,Y_1]}$ based on data (i^j, y_1^j), $j \in V^A$ and (i_2^j, y_1^j), $j \in V^{\overline{AB}}$, therefore needs the EM algorithm which in turn requires in each iteration the computation of $\nu^j(i) = \text{pr}(I = i | i_2^j, y_1^j)$, $j \in V^{\overline{AB}}$, which is zero for $i_2 \neq i_2^j$.

6. Discussion

As demonstrated in Section 2 and 3 the calculation of the ML estimates in the presence of missing values typically requires the application of the computer–intensive EM algorithm. The resulting computational effort can heavily increase when the EM algorithm is applied in models of high complexity such as graphical models. The approach presented in this paper to reduce the computational effort is based on the idea of taking special missing patterns into account when computing the ML estimates. It has been shown that for a certain kind of pattern the decomposition of the graph into subgraphs allowing separate maximization is possible even with missing values and essentially simplifies the algorithm. In special cases, ML estimates can even be explicitly determined, i.e. avoiding the EM algorithm.

The general approach proposed in Section 4 may give additional hints to further simplifications. If for example in a pure graph the subgraph $G_{B\cup D}$ is complete there exist closed expressions for the ML estimates in $\mathcal{M}(G_{B\cup D})^D$ not only for the situation that the whole vector X_B is either missing or observed but also when the missing pattern in X_B is monotone as described by Little & Rubin (1987).

Furthermore, if the sets A and B are not connected at all, that is $D = \emptyset$, then separate maximization of the likelihoods in $\mathcal{M}(G_A)$ and $\mathcal{M}(G_B)$ is possible regardless of the missing pattern. Of course one or both may require the EM algorithm.

As we have seen from our results, most simplifications are derived from a decomposition of a graph, where such a decomposition is often not unique. In that case it

should be chosen according to the missing pattern in order to apply the results of Section 4 and to get further decompositions if possible.
Having this in mind, it is straightforward to use the procedure proposed in Section 4 in the situation of G being collapsible onto a subset $A \subset V$ when the vectors X_{B_k} are incompletely observed for $k = 1, \ldots, K$ where $B_1, \ldots, B_K$ are the connected components of $B = V \backslash A$ since $(V \backslash \mathrm{cl}(B_k), B_k, \mathrm{bd}(B_k))$ is a decomposition of G for every $k = 1, \ldots, K$. It is intuitively clear that it can also be applied to decomposable graphs where the suitable missing patterns may even be more general.
Finally, it should be pointed out that another important situation where special missing patterns are worth to be taken into account is that of a chain graph. Here, the joint distribution is specified by conditional distributions each constituting a CG regression (Lauritzen & Wermuth, 1989). Missing patterns which considerably simplify the estimation task in these models are given when the "past" of a variable is always more observed than the variable itself since then regressions can be computed with complete covariable information.

Acknowledgments

We gratefully acknowledge the financial support of this paper by SFB 386 of the Deutsche Forschungsgesellschaft.

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM–algorithm. *J. R. Statist. Soc.* **B 39**, 1 – 38.

Edwards, D. (1996). *Introduction to graphical modelling.* New York: Springer.

Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18**, 790 – 805.

Frydenberg, M. & Edwards, D. (1989). A modified iterative proportional fitting algorithm for estimation in regular exponential families. *Comput. Statist. Data Anal.* **8**, 143 – 153.

Frydenberg, M. & Lauritzen, S.L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76**, 539 – 555.

Lauritzen, S.L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *J. Am. Statist. Ass.* **87**, 1098 – 1108.

Lauritzen, S.L. (1995). The EM algorithm for graphical association models with missing data. *Comput. Statist. Data Anal.* **19**, 191 – 201.

Lauritzen, S.L. (1996). *Graphical models.* Oxford: University Press.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Statist. Soc.* **B 50**, 157 – 224.

Lauritzen, S.L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31 – 57.

Leimer, H.G. (1989). Triangulated graph with marked vertices. *Ann. Discrete Math.* **41**, 311 – 324.

Little, J.A. & Rubin, D.B. (1987). *Data analysis with missing values.* New York: Wiley.

Little, J.A. & Schluchter, M.D. (1985). Maximum likelihood estimation from mixed continuous and categorical data with missing values. *Biometrika* **72**, 497 – 512.

Rubin, D.B. (1974). Inference and missing data. *Biometrika* **63**, 581 – 592.