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, didelez@stat.uni-muenchen.de

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). The approach via the EM algorithm of Little & Schluchter (1985) for the saturated case is expanded to cover the special restrictions in graphical models. A more efficient way to compute the E–step is indicated. 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 often of interest in contexts in which missing values are likely to occur and appropriate estimation procedures have to be found for estimating the parameters of interest. We focus here on maximum likelihood estimation in mixed graphical interaction models assuming a conditional Gaussian distribution, where maximum likelihood 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 § 2 we give a short introduction to graphical interaction models with conditional Gaussian distributions. The application of the EM algorithm for calculating the maximum likelihood estimates when the missing values occur at random is discussed in § 3. Since computational effort can be quite high, § 4 emphasises how one can simplify the algorithm given special missing patterns. An example illustrates the possible reduction in computational effort.

2. Graphical Models and Maximum Likelihood Estimation with complete data

We briefly introduce the graphical interaction models of interest using the terminology established by Lauritzen & Wermuth (1989). Consider a random vector $X_V = (Y^\top, I^\top)^\top$, where Y is a vector of R continuous variables with realisations $y \in I\!R^R$ and I is a vector of Q discrete variables with $\mathcal{I}$ denoting the set of possible realisations i. The index set V is divided into disjoint sets $V = \Gamma \cup \Delta$, $\Gamma \cap \Delta = \emptyset$, where Δ is the index set of the discrete components and Γ that of the continuous ones. The vector X_V is said to have a conditional Gaussian distribution if the density function $f(x_V)$ is given by the product of the discrete marginal probability $\mathrm{pr}(I = i) = p(i) > 0$ and 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 $i \in \mathcal{I}$. With the transformations

$$h(i) = \Sigma(i)^{-1}\mu(i) \quad \text{and} \quad K(i) = \Sigma(i)^{-1}$$

we have the standard mixed characteristics $\{p(i), h(i), K(i) | i \in \mathcal{I}\}$. The graphical models considered here specify conditional independencies which can be represented by a graph and which result in restrictions on the parameters (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, where we only consider undirected graphs. The multivariate

distribution of X_V is called G–Markovian if it fulfils the so–called pairwise Markov property:

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

For conditional Gaussian distributions this is equivalent to the global Markov property (Lauritzen & Wermuth, 1989).

We will denote by $\mathcal{M}(G)$ the statistical model containing all G–Markovian conditional Gaussian distributions and let us distinguish $\mathcal{M}(G)_A$ as the set of A–marginals from $\mathcal{M}(G_A)$ as the set of all G_A–Markovian conditional Gaussian distributions with $G_A = (A, E_A)$ and $E_A = E \cap (A \times A)$ for any $A \subset V$. In addition, let $\mathcal{M}(G)^A$ be the set of conditional G–Markovian conditional Gaussian distributions conditioning on the variables X_A.

Given a random sample $X_V^1, \ldots, X_V^N$ of independent and identically distributed random vectors from $\mathcal{M}(G)$ the set of joint distributions constitutes an exponential family. 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 $SS(i_d)_{r,s} = \sum_{j\in\mathcal{J}(i_d)} Y_r^j Y_s^j$, $d \in \mathcal{C}_\Delta(r,s)$,

where χ is the indicator function and $\mathcal{J}(i_d) = \{j \in \{1,\ldots,N\} | i_d^j = i_d\}$. The maximum likelihood estimates are given by the usual equation system. Sufficient conditions to guarantee the existence of the maximum likelihood estimates can be given but they are only necessary in the decomposable case (Lauritzen, 1996) which will be considered next.

A decomposition of a marked graph G is a partition (A,B,C) of V with (a) C

separates A and B, (b) C is complete and (c) $C \subseteq \Delta$ or $B \subseteq \Gamma$. A graph is decomposable if it is complete or if there exists a decomposition (A, B, C) with A and B both nonempty into decomposable subgraphs $G_{A\cup C}$ and $G_{B\cup C}$. Given such a decomposition (A, B, C) of G the graph is collapsible on to $A \cup C$, that is $\mathcal{M}(G)_{A\cup C} = \mathcal{M}(G_{A\cup C})$ (Frydenberg, 1990). The loglikelihood

$$L(\theta|x) = \sum_{j=1}^{N} \log f(x^j_{A\cup C}|\theta_{A\cup C}) + \sum_{j=1}^{N} \log f(x^j_B|x^j_C; \theta_{B|C}), \tag{1}$$

can then be maximised by separately maximising the two sums. It follows from the central result of Frydenberg & Lauritzen (1989, Proposition 4) that the first sum is maximised by the maximum likelihood estimate in $\mathcal{M}(G_{A\cup C})$ based upon data $(x^1_{A\cup C}, \ldots, x^N_{A\cup C})$ and the second by the maximum likelihood estimate in the regression model $\mathcal{M}(G_{B\cup C})^C$ based upon data $(x^1_{B\cup C}, \ldots, x^N_{B\cup C})$. The estimation in $\mathcal{M}(G_{B\cup C})^C$ in turn is given by the estimates in $\mathcal{M}(G_{B\cup C})$ and $\mathcal{M}(G_C)$. In addition, Frydenberg & Lauritzen (1989) show that closed expressions of the maximum likelihood estimates exist for decomposable graphs. In general iterative procedures are needed to calculate the maximum likelihood estimates (Frydenberg & Edwards, 1989).

3. Application of the EM algorithm

Incomplete data are modelled by dividing each observation vector into its observed and missing components, i.e. $X_V = (X^\top_{\text{obs}}, X^\top_{\text{mis}})^\top$. In the following we assume that for every entity at least one component of X_V can be observed. In addition we assume missingness at random (MAR) in the sense of Rubin (1974). 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 obtain the maximum likelihood estimates without any further knowledge about the missing mechanism. Their calculation requires maximisation of the likelihood of the observed variables. A general tool for handling this sometimes tedious task is the

EM algorithm (Dempster, Laird & Rubin, 1977) which is easy to apply when the considered model is an exponential family: the E–step calculates the expected sufficient statistics given the observed data and the current estimates of the parameters, and the M–step determines the new estimates using the conditional expectations of the sufficient statistics as if they were the observed. Thus, the M–step can be performed in the same way as for complete–data maximum likelihood estimation. For mixed interaction models with conditional Gaussian distribution the E–step has to calculate

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

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

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

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

These can be obtained by appropriate summation over the conditional expectations of the sufficient statistics in the saturated model (Little & Schluchter, 1985; Edwards, 1996). For example we have

$$\text{pr}(I_d = i_d|x_{\text{obs}}) = \sum_{i' \in \mathcal{I}: i'_d = i_d} \text{pr}(I = i'|x_{\text{obs}}).$$

To compute this let $(\mu(i)_{\text{obs}}, \Sigma(i)_{\text{obs}})$ be the parameters of the marginal distribution of Y_{obs} given $I = i$ and let $\mathcal{S} = \{(i_{\text{obs}}, i_{\text{mis}}) | i_{\text{mis}} \in \mathcal{I}_{\text{mis}}\}$ be the set of cells the observation could lie in given the observed discrete components. Then

$$\nu(i) = \text{pr}(I = i|x_{\text{obs}}) = \frac{\exp \kappa(i)}{\sum_{s \in \mathcal{S}} \exp \kappa(s)}$$

with

$$\kappa(i) = y^\top_{\text{obs}} \Sigma(i)^{-1}_{\text{obs}} \mu(i)_{\text{obs}}$$
$$-\frac{1}{2} \left[y^\top_{\text{obs}} \Sigma(i)^{-1}_{\text{obs}} y_{\text{obs}} + \mu(i)^\top_{\text{obs}} \Sigma(i)^{-1}_{\text{obs}} \mu(i)_{\text{obs}} \right] + \log p(i).$$

This differs slightly from the formulae given by Little & Schluchter (1985) since, because of the non–homogeneity assumption the term $\frac{1}{2}y_{\text{obs}}^{\top}\Sigma(i)_{\text{obs}}^{-1}y_{\text{obs}}$ 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 have, for missing continuous components Y_r, Y_s,

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

where $\text{cov}(Y_{\{r,s\}}|y_{\text{obs}}, i)$ denotes the conditional covariance matrix of $Y_{\{r,s\}}$ with entries $\text{cov}(Y_r, Y_s|y_{\text{obs}}, i)$ as conditional covariance of Y_r and Y_s, and $\text{var}(Y_r|y_{\text{obs}}, i)$ and $\text{var}(Y_s|y_{\text{obs}}, i)$ 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 $c_{r,s}(i) = 0$.
The conditional expectations of the sufficient statistics given the observed data are now given as follows:

$$E(N(i_d)|x_{\text{obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i'), \qquad d \in \mathcal{C}_\Delta, \tag{2}$$

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

$$E(S(i_d)_r|x_{\text{obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i') y_r^j(i'), \qquad d \in \mathcal{C}_\Delta(r),\ r \in \Gamma, \tag{3}$$

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

$$E(SS(i_d)_{r,s}|x_{\text{obs}}) = \sum_{j=1}^{N} \sum_{i' \in \mathcal{I}: i'_d = i_d} \nu^j(i')\{y_r^j(i') y_s^j(i') + c_{r,s}^j(i')\}, \qquad d \in \mathcal{C}_\Delta(r, s). \tag{4}$$

The E–step of the EM algorithm determines (2), (3) and (4) for the current parameter iterates. While (2) and (3) 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 (4) that the conditional covariances only have to be calculated for missing continuous components Y_r and Y_s with $(r, s) \in E$.

As noticed by Lauritzen (1995) the effort involved in the E–step can be considerable, especially when dealing with high dimensions. The following example makes it

clear that an acceleration is possible. To compute $E(N(i_d)|X_{\text{obs}})$, $d \in \mathcal{C}_\Delta$, we need $\text{pr}(I_d = i_d|x_{\text{obs}})$. If now the set of observed variables x_{obs} contains the boundary of d which is defined as $\text{bd}(d) = \{a \in V | \exists\, b \in d : (a,b) \in E \vee (b,a) \in E\}$, then it follows from the local Markov property that $\text{pr}(I_d = i_d|x_{\text{obs}}) = \text{pr}(I_d = i_d|x_{\text{bd}(d)})$ so that the computation depends on fewer variables. The corresponding simplification, however, is 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 only discrete variables but points out that the procedure can be generalised for mixed interaction models using the propagation scheme of Lauritzen (1992). Probabilistic expert systems specify the existing knowledge about association structures 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 factors. Computation can be done with unnormalised density functions and the Markov properties of the graph can be exploited in that they are reflected by the product structure of the joint density. For this it is necessary to form a junction tree, which is a special way of organising the cliques of the graph so that calculations can rely on operations only between neighbouring cliques. The operations in turn are done on conditional Gaussian potentials avoiding normalisation. For further details we refer to Lauritzen (1992).

4. Special Missing Patterns

In some situations, we can find simple factorised formulae for the marginal likelihood of the observed data. This is well known for monotone missing patterns and certain

underlying distributions as the multinomial and multivariate Gaussian (Little & Rubin, 1987) allowing a factorisation such that maximisation of each factor corresponds to a complete–data situation. In general this simplification only works for saturated models because maximising separately is often impossible when there are restrictions on the parameters. Much of the literature on graphical models, however, is concerned with simplifications of the estimation problem using the properties of decomposability and collapsibility of graphs leading to factorisations of the likelihood. For certain missing patterns the property of separate maximisation in these cases is preserved. We first describe the necessary general missing pattern and then discuss special cases yielding further simplifications.

Let (A, B, C) be a decomposition of the graph G. The missing pattern that will be of interest here obtains whenever X_C is 'more observed' than X_B. To describe this formally let obs(B) denote the observed components of a subvector X_B for any $B \subseteq V$. In a sample $X_V^1, \ldots, X_V^N$ the vector X_C is more observed than X_B if from obs$(B) \neq \emptyset$ it follows that obs$(C) = C$ for each observation. With $V^B \subseteq \{1, \ldots, N\}$, nonempty, being the index set of those observations where at least one component of X_B is known, the loglikelihood of the observed data is by analogy with (1) given by

$$L_{\text{obs}}(\theta | x_{\text{obs}}) = \sum_{j=1}^{N} \log f(x^j_{\text{obs}(A\cup C)} | \theta_{A\cup C}) + \sum_{j \in V^B} \log f(x^j_{\text{obs}(B)} | x^j_C; \theta_{B|C}),$$

where $f(x^j_{\text{obs}(A\cup C)} |\, \theta_{A\cup C})$ is the marginal density of the observed variables and $f(x^j_{\text{obs}(B)} |\, x^j_C; \theta_{B|C})$ that of $X_{\text{obs}(B)}$ given X_C. They are therefore given by marginalisation of the two sums in (1), preserving distinctiveness of the parameters. We therefore have the following result. The maximum likelihood estimates of the full model are transformations of the maximum likelihood estimates in

(i) $\mathcal{M}(G_{A\cup C})$ based on data $(x^j_{\text{obs}(A\cup C)};\ j \in \{1, \ldots, N\})$,

(ii) $\mathcal{M}(G_{B\cup C})$ based on data $(x^j_{\text{obs}(B)\cup C};\ j \in V^B)$ and

(iii) $\mathcal{M}(G_C)$ based on data $(x^j_C;\ j \in V^B)$.

For a more formal proof see Geng et al. (1997). Note that for the considered missing pattern (iii) always corresponds to a complete–data problem since X_C is always completely observed for $j \in V^B$, whereas in (i) and (ii) the EM algorithm may be required depending on the missing patterns within the vectors $X_{A\cup C}$ and $X_{B\cup C}$.

A special case not needing the EM algorithm in (ii) is when the whole vector X_B is either missing or observed. Then maximisation in $\mathcal{M}(G_{B\cup C})$ is based on $(x^j_{B\cup C};\ j \in V^B)$. Another easily handled situation obtains when the patterns in $A \cup C$ and $B \cup C$ are such that they again allow for a decomposition with corresponding factorisation. In the decomposable case with a so–called SD–ordering of the cliques (Leimer, 1989) the corresponding missing pattern is given when the separators are more observed than the separated sets. Further decomposition, however, will be impossible when A and B are complete. In this case, monotone missing patterns could lead to explicit maximum likelihood estimates. If for example the subgraph $G_{B\cup C}$ is complete and $\mathcal{M}(G_{B\cup C})$ corresponds to either a loglinear or a multivariate normal model there exist closed expressions for the maximum likelihood estimates in $\mathcal{M}(G_{B\cup C})^C$ 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. We can then apply the procedure described by Little & Rubin (1987).

If we have a symmetric decomposition which means that the sets A and B are interchangeable, as in the case of $\Gamma = \emptyset$ or $\Delta = \emptyset$, i.e. so–called pure graphs, or when C contains only discrete variables, the graph is collapsible on to C. We can then replace (i) by separate maximisation in $\mathcal{M}(G_{A\cup C})^C$ and $\mathcal{M}(G_C)$. This is preserved for a missing pattern where X_C is more observed not only than X_B but also than X_A so that maximisation in $\mathcal{M}(G_{A\cup C})^C$ is based on data $(x^j_{\mathrm{obs}(A)\cup C};\ j \in V^A)$, where V^A denotes the analogous set to V^B, and in $\mathcal{M}(G_C)$ on $(x^j_{\mathrm{obs}(C)};\ j \in \{1,\ldots,N\})$. The first of these two estimation tasks corresponds to (ii) and (iii) in the above result and therefore yields the same simplifications.

Furthermore, if the sets A and B are not connected at all, that is $C = \emptyset$, then separate maximisation 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.

Note that a decomposition is often not unique. In that case it should be chosen according to the missing pattern in order to apply the above results and to create further decompositions if possible. The procedure can also be applied when G is collapsible on to 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 $k = 1, \ldots, K$, where $\mathrm{cl}(B_k) = B_k \cup \mathrm{bd}(B_k)$ is the closure of the set B_k.

5. Example

Following Frydenberg & Lauritzen (1989) we consider the following graph: $G = (V, E)$ 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.

(Figure 1 about here)

With $A = \{I_1\}$, $B = \{Y_2\}$ and $C = \{I_2, Y_1\}$ we have a decomposition and since A and B are complete the graph is decomposable. 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 $(i_1, i_2) \in \mathcal{I}$. With complete data there exist explicit maximum likelihood estimates as given in Frydenberg & Lauritzen (1989). Let us now assume that Y_2 is incompletely observed. Each E–step of the EM algorithm would require the computation of

$$\begin{aligned} E(Y_2 | i^j, y_1^j) &= \mu(i^j)_2 + \frac{\sigma(i^j)_{12}}{\sigma(i^j)_1}\{y_1^j - \mu(i^j)_1\}, \\ \mathrm{var}(Y_2 | i^j, y_1^j) &= \sigma(i^j)_2 - \frac{\sigma(i^j)_{12}}{\sigma(i^j)_1}, \end{aligned}$$

for each incomplete observation $j \in V \backslash V^B$ and for the current parameter iterates. The EM algorithm is not complicated for this missing situation but can be avoided since we have the special missing pattern of § 4. In this special case, we have the following estimates for the standard mixed characteristics in the submodel $\mathcal{M}(G_{I_1,I_2,Y_1})$ estimated from all observations since the variables I_1, I_2 and Y_1 are always observed:

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

where ssd denotes the sum of squares of deviations from the mean. The estimates in $\mathcal{M}(G_{I_2,Y_1,Y_2})$ and $\mathcal{M}(G_{I_2,Y_1})$ 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 *. We then have

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

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

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

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 situation that 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 database 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 based on data (i^j, y_1^j) where I_1 is observed, and on (i_2^j, y_1^j) where I_1 is missing, therefore needs the EM algorithm which in turn requires in each iteration the computation of $\nu^j(i) = \text{pr}(I = i | i_2^j, y_1^j)$.

6. Discussion

For more general missing patterns it has been shown by Geng et al. (1997) that the EM algorithm can be considerably accelerated by imputing in the E–step only the minimal necessary data to yield a missing pattern as described in § 4. It is obvious that in this situation the discussed special cases may reduce the amount of missing data that needs to be imputed.

Finally, note that another important situation where special missing patterns are worth taking into account is that of a chain graph. Here, the joint distribution is specified by conditional distributions each constituting a conditional Gaussian regression (Lauritzen & Wermuth, 1989). The estimation task in these models is greatly simplified when the 'past' of a variable is always more observed than the variable itself.

Acknowledgement

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

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM–algorithm (with Discussion). *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 – 53.

FRYDENBERG, M. & LAURITZEN, S.L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* 76, 539 – 55.

GENG, Z., WAN, K., TAO, F. & GUO, J.H. (1997). Decomposition of mixed graphical models with missing data (Invited Paper). *Proc. Int. Symp. Contemp. Mult. Anal. Applic.*, 49 – 54.

LAURITZEN, S.L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *J. Am. Statist. Assoc.* 87, 1098 – 108.

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: 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 – 24.

LITTLE, R.J.A. & RUBIN, D.B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Little, R.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 – 92.

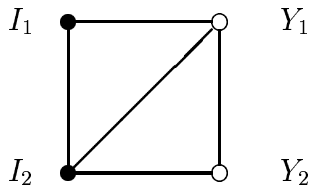

FIG. 1: A decomposable marked graph.