The ME Algorithm for Maximizing a Conditional Likelihood Function*

David Edwards
Novo Nordisk A/S

Steffen L. Lauritzen
Aalborg University

October 20, 1999

Abstract

This note describes an algorithm for maximization of a conditional likelihood function in the situation where the corresponding unconditional likelihood function can be more easily maximized. The algorithm is dual to the EM algorithm in the sense that the parameters rather than the data are augmented and that the conditional rather than the marginal likelihood function is maximized.

In exponential families the algorithm takes a particular simple form and the specific computations involved in the steps of the algorithm are identical to computations in the EM algorithm. To reflect the structure of the algorithm and the above-mentioned duality, we have chosen to refer to the algorithm as the ME algorithm.

The algorithm applies to mixed graphical chain models (Lauritzen and Wermuth 1989) and their generalizations (Edwards 1990), and it was developed with these as motivation, but we believe it to have potential applications beyond these.

The algorithm has been implemented in the most recent version of the MIM software (Edwards 1995).

Key words: CG distributions, conditional inference, graphical models, logistic regression.

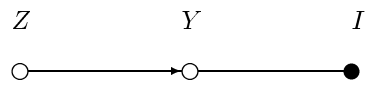
1 Introduction

Graphical chain models were introduced by Lauritzen and Wermuth (1989) and extended by Edwards (1990). Their potential for applications has been discussed by Wermuth and Lauritzen (1990), Cox and Wermuth (1996), and others. Their practical use has been limited by the absence of an estimation algorithm which can be easily implemented in general software for fitting

*This is Research Report R-99-2015, Department of Mathematical Sciences, Aalborg University.

and analysing graphical models. This note provides such an algorithm, thus meeting the challenge given on p. 219 of Lauritzen (1996).

The building blocks in graphical chain models and their extensions are based on so-called CG-regressions which describe the distribution of multiple discrete and continuous response variables given multiple discrete and continuous explanatory variables. They are derived by conditioning on the explanatory variables, in a corresponding family of joint models, the CG-distribution models, where a general algorithm for estimation has been described by Frydenberg and Edwards (1989) and implemented in the MIM software (Edwards 1995). In the special case where the explanatory variables form a cut (Barndorff-Nielsen 1978, p. 50) in the CG-distribution models, the CG-regression models can be fitted by piecing together suitable estimates obtained in the CG-distribution models as described in Proposition 6.33 of Lauritzen (1996). In particular this holds when all variables are discrete or all variables are continuous. The simplest case for which no general algorithm was available has the graph given below.



The structure of our algorithm is quite general and applies to maximizing any likelihood function obtained by conditioning on some observed variables, provided maximization can be performed in the unconditional model and the relevant conditional expectation can be calculated.

It resembles the EM algorithm (Dempster *et al.* 1977) by alternating between maximization of a function related to the true likelihood function (an M-step), and computation of a conditional expectation (an E-step), but it is different from this algorithm and its variants (Meng and van Dyk 1997) by being applied to the complete data case and by augmenting the parameters rather than the data. The idea of parameter extension is also exploited in Liu *et al.* (1998), but with a different purpose.

The algorithm is described in general terms in Section 2. It is most directly applicable to the exponential family case and this specialization is described in Section 3. Section 4.1 illustrates the algorithm by its application to linear logistic regression, a simple instance of the CG-regression models, which form the main motivation for the algorithm. Examples of the latter are discussed in Section 4.2.

2 The general structure of the algorithm

We consider a vector $Y = (Y_1, \ldots, Y_n)$ of response variables and a vector $X = (X_1, \ldots, X_n)$ of corresponding explanatory variables. Both the response and explanatory variables may themselves be multivariate.

We introduce the following notation for the unconditional, marginal and conditional log-likelihood functions

$$\begin{aligned} l(\theta) &= \log f(x, y \mid \theta) \\ l_x(\theta) &= \log f(x \mid \theta) \\ l^x(\theta) &= \log f(y \mid x, \theta). \end{aligned}$$

The marginal and conditional likelihood functions will typically be over-parametrized and most often only depend on a part of the parameter θ. Note that our object of interest l^x satisfies the relation $l^x = l - l_x$.

As with the EM algorithm, the computations of the algorithm are made in two separate steps. In one step we form the function

$$q(\theta) = q(\theta \mid \theta') = l(\theta) - \theta^\top \dot{l}_x(\theta'), \tag{1}$$

where θ' is a fixed value of the parameter θ and we have used a dot to denote differentiation with respect to θ. This function is an approximation to the conditional log-likelihood function, obtained by linearizing the marginal log-likelihood l_x, and then omitting terms constant with respect to θ. It involves the gradient $\dot{l}_x$ of the marginal log-likelihood function. Under regularity conditions of usual type, the gradient can be calculated as the conditional expectation of the score statistic in the unconditional model:

$$\dot{l}_x(\theta) = \mathbf{E}_\theta\{\dot{l}(\theta) \mid x\}, \tag{2}$$

because we have that

$$\dot{l}_x(\theta) = \mathbf{E}_\theta\{\dot{l}_x(\theta) \mid x\} = \mathbf{E}_\theta\{\dot{l}(\theta) - \dot{l}^x(\theta) \mid x\} = \mathbf{E}_\theta\{\dot{l}(\theta) \mid x\} - 0,$$

since the expected score statistic in the conditional model is equal to zero. As (2) involves the calculation of a conditional expectation, we refer to this as the *E-step.* Note that the computation in the E-step of the EM algorithm involves the conditional expectation of l rather than $\dot{l}$.

Note also that the functions q and the conditional likelihood function l^x have the same gradient at θ'

$$\dot{q}(\theta') = \dot{l}(\theta') - \dot{l}_x(\theta') = \dot{l}^x(\theta').$$

The M-step of the algorithm involves maximization of the function q defined above in (1)

$$\theta^* = \arg\max_\theta q(\theta \mid \theta').$$

The iteration is initiated by choosing θ_0 to maximize the unconditional likelihood function. The iteration therefore takes the form

$$\theta_0 = \arg\max_\theta l(\theta); \quad \theta_{n+1} = \arg\max_\theta q(\theta \mid \theta_n).$$

A GME variation of the algorithm just increases q instead of maximizing it.

In the special case where x induces a cut (Barndorff-Nielsen 1978, p. 50) no iteration is needed at all. Because then we have $\theta = (\psi, \eta)$ with ψ and η variation independent and

$$l(\theta) = l^x(\psi) + l_x(\eta),$$

implying that $\theta_0 = (\psi_0, \eta_0)$ satisfies

$$\psi_0 = \arg\max_{\psi} l^x, \quad \eta_0 = \arg\max_{\eta} l_x,$$

and thus $\dot{l}_x(\theta_0) = 0$, leading to $q = l$.

An important property of the EM algorithm is that the function to be maximized increases at each iteration. Unfortunately the situation here is not quite so simple. If we let

$$j_x(\theta) = -\ddot{l}_x(\theta)$$

be the observed marginal information about θ, we can expand an additional term of the marginal log-likelihood and get

$$\begin{aligned} l^x(\theta) &= l(\theta) - l_x(\theta') - (\theta - \theta')^\top \dot{l}_x(\theta') + (\theta - \theta')^\top j_x(\theta'')(\theta - \theta')/2 \\ &= q(\theta) - l_x(\theta') + \theta'^\top \dot{l}_x(\theta') + (\theta - \theta')^\top j_x(\theta'')(\theta - \theta')/2, \end{aligned}$$

for some θ'' between θ' and θ. Thus, if observed marginal information $j_x(\theta'')$ is non-negative we get

$$l^x(\theta^*) - l^x(\theta') = q(\theta^*) - q(\theta') + (\theta^* - \theta')^\top j_x(\theta'')(\theta^* - \theta')/2 \geq 0$$

such that in this case, the conditional likelihood will increase in each M-step of the algorithm. However it may happen — although rarely so — that the observed information has large negative eigenvalues and thus the conditional likelihood may in principle decrease. The M-step must then be modified by a line search. If the directional derivative of l^x towards θ^* is positive there will be a θ'' somewhere between θ' and θ^* satisfying

$$l^x(\theta'') > l^x(\theta').$$

Since q and l^x have the same gradient, this holds if $\dot{q}(\theta')^\top(\theta^* - \theta') > 0$. Else a full line search must be performed in the gradient direction.

We will generally assume that the M-step is executed with a line search modification, so that each iteration has

$$l^x(\theta_{n+1}) > l^x(\theta_n).$$

3 Exponential families

The prime application for the ME algorithm is the case of a steep exponential family (Barndorff-Nielsen 1978, p. 117). More precisely we assume that

$$l(\theta) = \alpha^\top u(x,y) + \beta^\top v(x) - \psi(\alpha,\beta)$$

and

$$l^x(\theta) = \alpha^\top u(x,y) - \psi^x(\alpha)$$

where

$$\psi(\alpha,\beta) = \log \int \exp\{\alpha^\top u(x,y) + \beta^\top v(x)\}\mu(dy \mid x)\mu(dx),$$

and

$$\psi^x(\alpha) = \log \int \exp\{\alpha^\top u(x,y)\}\mu(dy \mid x).$$

The parameters (α,β) are assumed to vary in the convex and open set

$$D = \operatorname{int}\{(\alpha,\beta) \mid \psi(\alpha,\beta) < \infty\}$$

and the steepness assumption implies that the gradient of ψ tends to ∞ when (α,β) approaches the boundary of D.

Let $U = u(X,Y)$ and $V = v(X)$. It is convenient to introduce the mixed parametrization $\theta = (\alpha,\eta)$, where

$$\eta = \mathbf{E}_\theta\{V\} = \frac{\partial}{\partial\beta}\psi(\alpha,\beta).$$

The mixed parameters are variation independent (Barndorff-Nielsen 1978, p. 122), so x induces a cut if and only if the distribution of X only depends on θ through η.

If the observed (u,v) is in the interior of the convex support of (U,V), the unconditional likelihood function has a unique maximum determined by

$$\eta = v, \quad \mathbf{E}_\theta\{U\} = u,$$

whereas the conditional likelihood function has its unique maximum at the point where

$$\mathbf{E}_\theta\{U \mid x\} = u. \tag{3}$$

The E-step of the algorithm takes a particularly simple form. We find for any $\theta = (\alpha,\eta)$ with $\eta = v$ that

$$\begin{aligned}
\mathbf{E}_\theta\left\{\frac{\partial}{\partial\eta}l(\theta) \mid x\right\} &= \mathbf{E}_\theta\left\{V^\top\frac{\partial\beta}{\partial\eta} - \frac{\partial}{\partial\beta}\psi(\alpha,\beta)^\top\frac{\partial\beta}{\partial\eta} \mid x\right\} \\
&= (v-\eta)^\top\frac{\partial\beta}{\partial\eta} = 0.
\end{aligned}$$

Also

$$\mathbf{E}_\theta\left\{\frac{\partial}{\partial\alpha} l(\theta)\,|\,x\right\} = \mathbf{E}_\theta\{U\,|\,x\} - \frac{\partial}{\partial\alpha}\psi(\alpha,\beta) = \mathbf{E}_\theta\{U\,|\,x\} - \mathbf{E}_\theta\{U\}$$

and thus

$$\theta^\top \dot{l}_x(\theta') = \alpha^\top\delta + \eta^\top 0 = \alpha^\top\delta,$$

where

$$\delta = \mathbf{E}_{\theta'}\{U\,|\,x\} - \mathbf{E}_{\theta'}\{U\}.$$

Hence the E-step involves computation of the conditional and unconditional expectations of U and forming the function

$$q(\theta) = l(\theta) - \alpha^\top\delta.$$

The direct iteration therefore takes the following simple form

$$\theta_0 = \hat{\theta}(u,v); \quad u_{n+1} = u_n + u - \mathbf{E}_{\theta_n}\{U\,|\,x\}; \quad \theta_{n+1} = \hat{\theta}(u_{n+1},v).$$

If we introduce the mean value parameter

$$\tau = \mathbf{E}_\theta\{U\} = \frac{\partial}{\partial\alpha}\psi(\alpha,\beta),$$

the iteration can be expressed as briefly as

$$\tau_0 = u; \quad \tau_{n+1} = \tau_n + u - \mathbf{E}_{\theta_n}\{U\,|\,x\} \tag{4}$$

In general the iteration needs to be modified with line search, both because the conditional likelihood function may decrease, but also because u_{n+1} may be outside the convex support of U, in which case $\hat{\theta}(u_{n+1},v)$ does not exist. The modification can here be performed by instead letting

$$\tau_{n+1} = \tau_n + \lambda\,(u - \mathbf{E}_{\theta_n}\{U\,|\,x\})$$

in (4), where λ between 0 and 1 is found such that $l^x(\theta_{n+1}) > l^x(\theta_n)$. That such a λ exists, follows from the calculation below, showing that the directional derivative with respect to τ of the conditional log-likelihood function in the direction of u_{n+1} is positive: We have that

$$\begin{aligned}
\frac{\partial}{\partial\tau} l^x(\theta_n) &= \left(\frac{\partial\alpha}{\partial\tau}\right)^\top \left(u - \frac{\partial}{\partial\alpha}\psi^x(\alpha)\right) \\
&= \left(\frac{\partial\alpha}{\partial\tau}\right)^\top (u - \mathbf{E}_{\theta_n}\{U\,|\,x\}) \\
&= v(\theta_n)^{-1}_{11}\,(u - \mathbf{E}_{\theta_n}\{U\,|\,x\}),
\end{aligned}$$

where

$$v(\theta)^{-1}_{11} = \frac{\partial\alpha}{\partial\tau}$$

is the inverse of the covariance matrix of U. As this is positive definite, we get for the directional derivative

$$\begin{aligned}\left(\frac{\partial}{\partial\tau}l^x\right)^\top (u_{n+1}-u_n) &= \left(\frac{\partial}{\partial\tau}l^x\right)^\top (u-\mathbf{E}_{\theta_n}\{U\,|\,x\}) \\ &= (u-\mathbf{E}_{\theta_n}\{U\,|\,x\})^\top v(\theta_n)_{11}^{-1}\,(u-\mathbf{E}_{\theta_n}\{U\,|\,x\}) > 0.\end{aligned}$$

The essential computational task in the E-step is the computation of the conditional expectation $\mathbf{E}_{\theta_n}\{U\,|\,x\}$. This is identical to the computation in the E-step of the EM algorithm in the exponential family case. Thus methods used to speed up this computation (Lauritzen 1995; Geng *et al.* 1996, 2000; Didelez and Pigeot 1998) can also be exploited for the ME algorithm without further modification.

Any fixed point of the iteration must satisfy (3), so the conditional maximum likelihood estimate is the unique fixed point. As the conditional likelihood increases at every step, the algorithm must converge to the desired conditional maximum likelihood estimate.

4 Examples

4.1 Logistic regression

To illustrate the algorithm we consider the simple logistic regression model

$$P_{(\alpha,\beta)}(I=1\,|\,X=x) = \exp(\alpha+\beta x)/\{1+\exp(\alpha+\beta x)\}, \tag{5}$$

where I is binary $(0,1)$ and X a real-valued variable. We wish to find the maximum likelihood estimates of α and β. To do this we imbed the model in a joint model for I and X in which I is binomial with probability $\{p_i\}_{i=0,1}$ and for given $I=i$, $X\sim\mathcal{N}(\mu_i,\sigma^2)$. Given X, the distribution of I is of the form (5) with parameters

$$\begin{aligned}\alpha &= \ln(p_1/p_0)-(\mu_1-\mu_0)/2\sigma^2 \\ \beta &= (\mu_1-\mu_0)/\sigma^2\end{aligned}$$

Given a sample of the form $(i^v,x^v), v=1\ldots N$, the minimal canonical statistics under the joint model can be written (u,v), where $u=(n_0,t_0,t_1)$ and $v=(ss)$ with $n_0=\#\{v:i^v=0\}$, $t_j=\sum_{v:i^v=j}x^v$ for $j=0,1$, and $ss=\sum_{v=1\ldots N}(x^v)^2$. The algorithm requires computation of $\mathbf{E}_\theta\{U\,|\,x\}$: this is given by the expressions

$$\begin{aligned}\mathbf{E}_\theta(N_0\,|\,x) &= \sum_{v=1\ldots N} P_\theta(I=0\,|\,X=x^v) \\ \mathbf{E}_\theta(T_0\,|\,x) &= \sum_{v=1\ldots N} P_\theta(I=0\,|\,X=x^v)x^v \\ \mathbf{E}_\theta(T_1\,|\,x) &= \sum_{v=1\ldots N} P_\theta(I=1\,|\,X=x^v)x^v\end{aligned}$$

The maximum likelihood estimates $\hat{\theta}(u,v)$ are given by $\hat{p}_0 = n_0/N$, $\hat{\mu}_0 = t_0/n_0$, $\hat{\mu}_1 = t_1/(N-n_0)$, and $\hat{\sigma}^2 = (ss - n_0\hat{\mu}_0^2 - (N-n_0)\hat{\mu}_1^2)/N$.

For a numerical example, we use the data described in Jensen *et al.* (1991) consisting of 4 replicates of $(0,-1)$, 11 of $(1,-1)$, 23 of $(0,1)$, 7 of $(1,1)$, 12 of $(0,2)$ and 3 of $(1,2)$, that is, 60 observations in all. Here $(u,v) = (39,43,2,105)$. The progress of the algorithm is shown in Table 1. The maximum absolute difference between u and $\mathbf{E}_{\theta_n}\{U \mid x\}$ is shown as d_n. The algorithm stops when $d_n < 10^{-5}$.

n	u_n			$-2l^x(\theta_n)$	d_n
0	39.000000	43.000000	2.000000	66.173161	2.037482
1	38.624593	40.962518	4.037482	65.893845	0.562286
2	38.750190	41.524803	3.475197	65.875077	0.144482
3	38.721279	41.380322	3.619678	65.873788	0.038134
4	38.729209	41.418456	3.581544	65.873699	0.009989
5	38.727158	41.408467	3.591533	65.873693	0.002623
6	38.727698	41.411089	3.588911	65.873692	0.000688
7	38.727557	41.410401	3.589599	65.873692	0.000181
8	38.727594	41.410582	3.589418	65.873692	0.000047
9	38.727584	41.410535	3.589465	65.873692	0.000012
10	38.727587	41.410547	3.589453	65.873692	0.000003

Table 1: Progress of the ME algorithm in the logistic regression example.

4.2 CG-regression models

In this section we illustrate the use of the algorithm with the class of CG-regression models. They include, as simplest case, the simple logistic regression model of the previous example and, as mentioned in the introduction, they are derived by conditioning on the explanatory variables, in a corresponding family of joint models, the CG-distribution models.

As in the previous example, the minimal canonical statistics for a general CG-distribution model consist of a set of marginal cell counts, variate totals and variate sums of squares and products. A subset of these (those that are still stochastic after conditioning) comprise the minimal canonical statistics for the corresponding conditional model. The ME algorithm proceeds as before by iteratively incrementing the latter set with the difference between the observed statistics and their conditional expectations under the current parameter estimate, and then re-estimating using the incremented quantities.

We illustrate this using artificial data consisting of 28 observations of two binary variables (denoted I and J), and two real-valued variables (denoted Y and Z), shown in Table 2.

I	J	(y^v, z^v)						
0	0	3 2	3 3	4 7	4 8	3 2	5 4	6 8
0	1	3 3	3 4	2 4	1 4	2 7	3 5	4 5
1	0	4 8	4 7	3 8	3 9	4 3	5 8	2 4
1	1	1 4	1 5	3 7	3 6	4 6	5 8	6 9

Table 2: The data used in the CG-regression examples.

Table 3 summarizes the progress of the algorithm when applied to several CG-regression models. The second column shows the formula of the augmented model, using the syntax described in Edwards (1995). The convergence criterion is based on the likelihood equations, i.e. that

$$\max\left\{\frac{|\delta n|}{\sqrt{n}}, \frac{|\delta t^\gamma|}{\sqrt{ss^{\gamma\gamma}}}, \frac{|\delta ss^{\gamma\eta}|}{\sqrt{ss^{\gamma\gamma}ss^{\eta\eta} + (ss^{\gamma\eta})^2}}\right\} < 10^{-5},$$

where n, t^γ, and $ss^{\gamma\eta}$ are the marginal cell counts, totals and sums of squares in the minimal canonical statistics, and δn, δt^γ and $\delta ss^{\gamma\eta}$ are the differences between these and their conditional expectations.

Line-search is here implemented as a crude step-halving procedure: that is, if the step

$$u_{n+1} = u_n + u - \mathbf{E}_{\theta_n}\{U \mid x\}; \quad \theta_{n+1} = \hat{\theta}(u_{n+1}, v),$$

does not lead to $l^x(\theta_{n+1}) > l^x(\theta_n)$ then steps of the form

$$u_{n+1} = u_n + \lambda(u - \mathbf{E}_{\theta_n}\{U \mid x\}); \quad \theta_{n+1} = \hat{\theta}(u_{n+1}, v),$$

for $\lambda = \frac{1}{2}, \frac{1}{4} \ldots$ are successively examined until one is found for which the conditional loglikelihood does increase.

Note that the first and fourth models induce the same CG-regression model. Convergence is slower in the latter case, and line-search is required. This illustrates a general point that is taken up in the discussion below. The fifth model is heterogeneous, and so the CG-regressions involve quadratic terms. Convergence is slow for this model also, and line-search is again required. Here the large dimension of the conditional model relative to the size of the dataset may be noted. The sixth model has mixed responses and explanatory variables. The seventh model illustrates that the algorithm converges after one iteration when the explanatory variables form a cut.

No.	Formula of augmented model	Explanatory variables	No. of cycles	Line-search necessary
1	$IJ/IJY, IJZ/YZ$	$\{Y, Z\}$	9	no
2	$IJ/JY, IJZ/YZ$	$\{Y, Z\}$	7	no
3	$IJ/IJY, JZ/YZ$	$\{Y, Z\}$	4	no
4	$IJ/IJY, IJZ/Y, Z$	$\{Y, Z\}$	27	yes
5	$IJ/IJY, IJZ/IJYZ$	$\{Y, Z\}$	85	yes
6	$IJ/IJY, JZ/YZ$	$\{I, Y\}$	17	no
7	$IJ/IJY, IJZ/Y, Z$	$\{I, J\}$	1	no

Table 3: Progress of the ME-algorithm with different CG-regression models.

5 Discussion

Although the algorithm described was developed with the chain graph models in mind, the structure of the algorithm promises potential application in quite general contexts. For any given family of models used for studying the effects of explanatory variables on responses, one just needs to augment the parameter space by equipping the explanatory variables with a suitable marginal distribution, so that the joint distribution becomes simple. Admittedly this can be a more difficult task than the corresponding data augmentation needed for the EM algorithm which has by now established its wide applicability.

We emphasize that when the augmented joint family can be chosen in different ways, the analysis in Section 2 indicates that one should choose this family to be as rich as possible, to prevent strongly negative observed information in the marginal distribution about the parameters of the conditional distribution. This point is well illustrated by the first and fourth CG-distribution models above, inducing the same CG-regression model, the first being richer and leading to faster convergence. Similarly it is not an advantage to assume σ^2 to be known in the logistic regression example, although this will generate the same logistic regression model. If the marginal distribution is sufficiently rich, the explanatory variables will form a cut, such as in the last model in Section 4.2, and the algorithm will converge after a single cycle.

Note that as the entire action of the algorithm happens in the mean value space, as captured in (4), the iteration will also converge to a valid estimate when (u, v) is on the boundary of the convex support, provided the exponential family is extended with weak limits as described in Barndorff-Nielsen (1978), p. 157. This is important for models involving sparse contingency tables.

Finally we mention that a hybrid of the ME and EM algorithm appears when there are data missing on the response variables. The iteration (4) should then be replaced by

$$\tau_{n+1} = \tau_n + \mathbf{E}_{\theta_n}\{U \mid x, y_{\text{obs}}\} - \mathbf{E}_{\theta_n}\{U \mid x\},$$

from an arbitrarily chosen initial value of τ. We refrain from discussing convergence properties of this hybrid algorithm.

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* John Wiley and Sons, New York.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies.* Chapman and Hall, London.

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

Didelez, V. and Pigeot, I. (1998). Maximum likelihood estimation in graphical models with missing values. *Biometrika*, **85**, 960–6.

Edwards, D. (1990). Hierarchical interaction models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 3–20 and 51–72.

Edwards, D. (1995). *Introduction to Graphical Modelling.* Springer-Verlag, New York.

Frydenberg, M. and Edwards, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Computational Statistics and Data Analysis*, **8**, 143–53.

Geng, Z., Asano, C., Ichimura, M., Tao, F., Wan, K., and Kuroda, M. (1996). Partial imputation method in the EM algorithm. In *Compstat 96*, (ed. A. Prat), pp. 259–63. Physica Verlag, Heidelberg, Germany.

Geng, Z., Wan, K., and Tao, F. (2000). Mixed graphical models with missing data and the partial imputation EM algorithm. *Scandinavian Journal of Statistics*, **27**, to appear.

Jensen, S. T., Johansen, S., and Lauritzen, S. L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, **78**, 867–77.

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

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

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

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, **85**, 755–70.

Meng, X.-L. and van Dyk, D. (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 511–67.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *Journal of the Royal Statistical Society, Series B*, **52**, 21–72.