The TM Algorithm for Maximising a Conditional Likelihood Function

BY DAVID EDWARDS

Novo Nordisk A/S, Novo Alle, DK 2880 Bagsværd, Denmark

ded@novonordisk.com

AND STEFFEN L. LAURITZEN

Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, DK-9220 Aalborg Ø, Denmark

steffen@math.auc.dk

SUMMARY

This article describes an algorithm for maximising a conditional likelihood function when the corresponding unconditional likelihood function is more easily maximised. The algorithm is similar to the EM algorithm but different as the parameters rather than the data are augmented and the conditional rather than the marginal likelihood function is maximised.

In exponential families the algorithm takes a particular simple form which is computationally very close to the EM algorithm. The algorithm alternates between a T-step which calculates a tilted version of the unconditional likelihood function and an M-step which maximises it.

The algorithm applies to mixed graphical chain models (Lauritzen & Wermuth, 1989) and their generalisations (Edwards, 1990), and was developed with these in mind, but it may have applications beyond these.

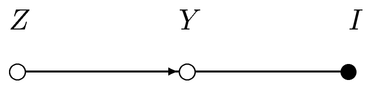
The algorithm has been implemented in the most recent version of the MIM software (Edwards, 2000), where it was named the ME algorithm. The name has been changed to avoid confusion with the algorithm described by Marschner (2001).

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

1 Introduction

Graphical chain models were introduced by Lauritzen & Wermuth (1989) and extended by Edwards (1990). Their potential for applications has been discussed by Wermuth & Lauritzen (1990), Cox & 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 and analysing graphical models. This article provides such an algorithm.

The building blocks in graphical chain models and their extensions are the so-called CG-regression models. These 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, for which a general estimation algorithm has been described by Frydenberg & Edwards (1989) and implemented in the MIM software. 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, where Z is a continuous explanatory variable, and (I, Y) a pair of response variables with I discrete and Y continuous.



Our algorithm is quite general and may be applied to maximise any likelihood function obtained by conditioning on some observed variables, provided maximisation 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 maximisation of a function related to the true likelihood function, i.e. an M-step, and computation of a conditional expectation to identify this function, but differs from the EM algorithm and its variants (Meng & 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 another purpose. Jebara & Pentland (1998) discuss a different algorithm for maximising conditional likelihood functions with emphasis on problems caused by missing data.

The algorithm is described in general terms in § 2. It is most directly applicable to the exponential family case and this specialisation is described in § 3, where also convergence properties are discussed. In § 4.1 the algorithm is illustrated by its application to linear logistic regression, a simple instance of the CG-regression models, which form the main motivation. Examples of CG-regression models are discussed in § 4.2 and further examples in § 4.3.

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-parameterised 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 tilted unconditional loglikelihood 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 loglikelihood function, obtained by linearising the marginal loglikelihood l_x, and then omitting terms constant with respect to θ. The tilt is determined by the gradient $\dot{l}_x$ of the marginal loglikelihood function. We refer to this step as the T-step of the algorithm.

Under regularity conditions of the usual type, the gradient can be calculated as the conditional expectation of the score statistic in the unconditional model:

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

because we have that

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

since the expected score statistic in the conditional model is equal to zero. This implies that the computation in the T-step is very similar to that in the E-step of the EM algorithm. Although the latter involves the conditional expectation of l rather than $\dot{l}$, this difference is immaterial as the M-step of the EM algorithm is typically made by solving the equation

$$\frac{\partial}{\partial\theta} E_{\theta'}\{l(\theta) \mid x\} = E_{\theta'}\{\dot{l}(\theta) \mid x\} = 0$$

and hence involves computation of essentially the same conditional expectation.

Note also that the tilted loglikelihood function q and the conditional loglikelihood 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 maximisation of the tilted loglikelihood function q defined above in (1):

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

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

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

so the M-step typically solves the equation

$$E_{\theta_n}\{\dot{l}(\theta_n) \mid x\} = \dot{l}(\theta), \tag{3}$$

in contrast to the EM algorithm which for the M-step solves $E_{\theta_n}\{\dot{l}(\theta) \mid x\} = 0$.

In the special case where x induces a cut (Barndorff-Nielsen, 1978, p. 50) no iteration is needed, 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 maximised 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 loglikelihood 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 the observed marginal information $j_x(\theta'')$ is nonnegative we get

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

so 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 then the conditional likelihood may decrease. The M-step must then be modified by a line search; this is discussed in more detail in § 3.2. The possible situations are illustrated in Fig. 1.

Insert Figure 1 about here.

3 Exponential families

3.1 General structure

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

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

where

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

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

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

and steepness means 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 parameterisation $\theta = (\alpha,\eta)$, where

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

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.

The steepness assumption implies that, 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

$$\eta = v, \quad E_\theta(U) = u,$$

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

$$E_\theta\left(U \,|\, x\right) = u. \tag{4}$$

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

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

Also

$$E_\theta \left\{ \frac{\partial}{\partial \alpha} l(\theta) \,\middle|\, x \right\} = E_\theta(U \mid x) - \frac{\partial}{\partial \alpha} \psi(\alpha, \beta) = E_\theta(U \mid x) - E_\theta(U)$$

and thus

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

where

$$\delta = E_{\theta'}(U \mid x) - E_{\theta'}(U).$$

Hence the T-step involves computation of the difference between 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:

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

If we introduce the mean value parameter

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

the iteration can be expressed as briefly as

$$\tau_0 = u; \quad \tau_{n+1} = \tau_n + u - E_{\theta_n}(U \mid x). \tag{5}$$

In general the iteration needs to be modified with line search, both because the conditional likelihood function may decrease and 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

$$\tau_{n+1} = \tau_n + \lambda_n \left\{ u - E_{\theta_n}(U \mid x) \right\}$$

in (5), where λ_n is a suitable value between 0 and 1; see § 3.2 below for details.

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

3.2 Convergence properties

To discuss convergence properties of line search modifications of the TM algorithm in the case of an exponential family we find the gradient of $l^x(\theta)$ with respect to τ as

$$\begin{aligned} \frac{\partial}{\partial \tau} l^x(\theta) &= \frac{\partial \alpha}{\partial \tau}^\top \left\{ u - \frac{\partial}{\partial \alpha} \psi^x(\alpha) \right\} \\ &= \frac{\partial \alpha}{\partial \tau}^\top \left\{ u - E_\theta(U \mid x) \right\}. \end{aligned}$$

Using the fact that $\partial\alpha/\partial\tau$ is equal to the inverse of the covariance matrix $\mathrm{var}_{\theta_n}(U)$ and letting $d_n = u - E_{\theta_n}(U \mid x)$, we find the directional derivative in the direction of $u_{n+1} = u_n + d_n$ as

$$l^x_{d_n}(\theta_n) = \left. \frac{\partial}{\partial t} l^x\{\theta(u_n + t d_n)\} \right|_{t=0} = d_n^\top \frac{\partial}{\partial \tau} l^x(\theta_n) = ||d_n||^2_{\theta_n},$$

where we have let $||x||^2_{\theta_n} = x^\top \mathrm{var}_{\theta_n}(U)^{-1} x$. Note that this derivative is strictly positive unless $d_n = 0$.

The aim of the line search is to choose λ_n with $0 < \lambda_n < 1$ in such a way that the sequence $\theta_n = \theta(\tau_n)$ determined by

$$\tau_{n+1} = \tau_n + \lambda_n d_n$$

is convergent, with the conditional maximum likelihood estimate $\hat{\theta}^x$ as its limit. A crude attempt would be to choose λ_n as the largest $\lambda \in \{2^{-j}\}_{j \geq 0}$ such that

$$l^x\{\theta(\tau_n + \lambda d_n)\} > l^x(\theta_n), \tag{6}$$

ensuring that the likelihood increases at every step. Since the directional derivative is positive, such a λ_n always exists. Although this line search will often suffice in practice, and has done so in all examples we have considered, it is not sufficient to ensure theoretical convergence, which in addition needs that $d_n \to 0$; see 14.1.4 of Ortega & Rheinboldt (1970, p. 475).

If λ_n defined above is bounded below by a constant for all n, an argument as in the proof of Proposition 2 of Frydenberg & Edwards (1989) yields that if the sequence of parameters converges, its limit is $\hat{\theta}^x$ as desired.

Generally, stronger forms of line search may be needed. One such version is the Goldstein–Armijo algorithm (Ortega & Rheinboldt, 1970, p. 491). This can for example be implemented as follows. Fix an arbitrary γ with $0 < \gamma < 1$ and define $\theta_n = \theta(\tau_n)$ by

$$\tau_{n+1} = \tau_n + \lambda_n^* d_n$$

where λ_n^* is chosen as the largest $\lambda \in \{2^{-j}\}_{j \geq 0}$ such that

$$l^x\{\theta(\tau_n + \lambda d_n)\} - l^x(\theta_n) \geq \gamma\lambda ||d_n||^2_{\theta_n}. \tag{7}$$

The following argument shows that λ_n^* always exists. Define λ_n as above in the crude algorithm. Consider the function

$$\phi(t) = \frac{l^x\{\theta(\tau_n + t d_n)\} - l^x(\theta_n)}{t\,||d_n||^2_{\theta_n}}.$$

If λ_n does not satisfy (7) we have $\phi(\lambda_n) < \gamma$. Application of l'Hôpital's rule yields that ϕ is continuous on the interval from 0 to λ_n and has $\phi(0) = 1$. Therefore ϕ must attain all values between γ and 1 on the interval $(0, \lambda_n)$ and λ_n^* must exist.

This choice of λ_n^* implies not only that the likelihood increases but also that it increases sufficiently at each step to ensure convergence of the algorithm. This follows from 14.2.15 of Ortega & Rheinboldt (1970), which yields

$$\lim_{n \to \infty} ||d_n||^2_{\theta_n} / ||d_n|| = 0, \tag{8}$$

where $||x||^2 = x^\top x$ denotes the usual Euclidean norm. But since for any bounded sequence θ_n there exist positive constants c and C so that

$$c||x||^2 \leq ||x||^2_{\theta_n} \leq C||x||^2,$$

(8) implies that $d_n \to 0$ and thus $\lim_{n \to \infty} \theta_n = \hat{\theta}^x$ as desired.

The disadvantage of the Goldstein–Armijo algorithm is that the asymptotic covariance matrix $\mathrm{var}_{\theta_n}(U)$ must be calculated at each step. The comparison principle (Ortega & Rheinboldt, 1970, p. 484) implies that choosing $\tilde{\lambda}_n$ by full line maximisation, i.e. such that

$$l^x\{\theta(u_n + \tilde{\lambda}_n d_n)\} = \sup_{t\in(0,\lambda_n)} l^x\{\theta(\tau_n + td_n)\}$$

also renders the algorithm $\tau_{n+1} = \tau_n + \tilde{\lambda}_n d_n$ convergent. Such full line maximisation can be performed with general numerical methods (Press et al., 1988, Ch. 10).

4 Examples

4.1 Logistic regression

To illustrate the algorithm we consider the simple logistic regression model

$$P_{(\alpha,\beta)}(I = 1 \,|\, X = x) = \exp(\alpha + \beta x)/\{1 + \exp(\alpha + \beta x)\}, \tag{9}$$

where I is binary $(0, 1)$ and X is 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 (9) with parameters

$$\alpha = \ln(p_1/p_0) - (\mu_0^2 - \mu_1^2)/2\sigma^2, \quad \beta = (\mu_1 - \mu_0)/\sigma^2.$$

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

$$\begin{aligned} E_\theta(N_0 \,|\, x) &= \sum_{k=1}^N P_\theta(I = 0 \,|\, X = x^k) \\ E_\theta(T_0 \,|\, x) &= \sum_{k=1}^N P_\theta(I = 0 \,|\, X = x^k)x^k \\ E_\theta(T_1 \,|\, x) &= \sum_{k=1}^N P_\theta(I = 1 \,|\, X = x^k)x^k. \end{aligned}$$

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

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 $E_{\theta_n}(U \mid x)$ is shown as d_n. The algorithm stops when $d_n < 10^{-5}$.

Insert Table 1 about here.

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 the 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 TM 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 by I and J, and two real-valued variables, denoted by Y and Z, shown in Table 2.

Insert Table 2 about here.

Table 3 summarises 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 (1990). The convergence criterion is based on the likelihood equations, i.e. that

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

where n, t^{γ} and $s^{\gamma\eta}$ are the marginal cell counts, totals and sums of squares in the minimal canonical statistics, and δn, δt^{γ} and $\delta s^{\gamma\eta}$ are the differences between these and their conditional expectations. Line search is here implemented using the crude step-halving procedure (6).

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 and seventh model have mixed responses and explanatory variables. The eighth model illustrates that the algorithm converges after one iteration when the explanatory variables form a cut.

Insert Table 3 about here.

4.3 Additional examples

In many cases it is not easy to find a suitable marginal distribution for the explanatory variables as required by the algorithm. However, we emphasise that the CG-regression models contain many standard models as special cases (Wermuth & Lauritzen, 1990), and also less standard models such as quadratic logistic regression, with

$$P_{(\alpha,\beta,\gamma)}(I=1 \mid X=x) = \exp(\alpha+\beta x+\gamma x^2)/\{1+\exp(\alpha+\beta x+\gamma x^2)\}$$

which is obtained by conditioning as in §4.1 without assuming variance homogeneity. Below we describe a few other instances of the algorithm for illustration.

Example 1: Poisson regression. Consider $Y_i, i = 1, \ldots, N$, independent and Poisson distributed with mean θx_i where $\theta > 0$ is unknown and $x_1, \ldots, x_N$ are fixed explanatory variables. The maximum likelihood estimator of θ can be explicitly found as $\hat{\theta}^x = u/v$, where $u = \sum y_i$ and $v = \sum x_i$ so the TM algorithm is less interesting in this case.

Still, for illustration we augment the model by assuming that X_i are exponentially distributed with mean $1/\theta$, leading to the unconditional loglikelihood function

$$l(\theta) = (u + N)\log\theta - 2v\theta,$$

where we have ignored a constant term. The gradient is equal to

$$\dot{l}(\theta) = (u + N)/\theta - 2v$$

and

$$\dot{l}_x(\theta) = E_\theta\{\dot{l}(\theta) \mid x\} = N/\theta - v,$$

so equation (3) yields the TM algorithm as

$$\theta_0 = (u + N)/(2v); \quad \theta_{n+1} = (u + N)/(v + N/\theta_n).$$

Example 2: von Mises–Fisher distribution. Let $Y_i, i = 1, \ldots, N$, be independent and distributed as the d-dimensional von Mises–Fisher distribution with density

$$f_d(y \mid \mu, \kappa) = a_d(\kappa)^{-1}\exp(\kappa\mu^\top y)$$

on the d-dimensional unit sphere $||y|| = 1$. Here

$$a_d(\kappa) = (2\pi)^{d/2} I_{d/2-1}(\kappa)\kappa^{1-d/2},$$

where $I_\nu(\kappa)$ is the modified Bessel function of the first kind and order ν.

As shown in Mardia (1975), the maximum likelihood estimator of (μ, κ) is given as the unique solution $(\hat{\mu}, \hat{\kappa})$ to the equations $\hat{\mu} = s/r$ and $A_d(\hat{\kappa}) = r/N$, where $s = \sum y_i$, $r = ||s||$ and A_d is the logarithmic derivative of a_d:

$$A_d(\kappa) = a_d(\kappa)'/a_d(\kappa) = I_{d/2}(\kappa)/I_{d/2-1}(\kappa).$$

It is well known and easily seen that this distribution can be obtained from the d-variate Gaussian distribution $Y \sim \mathcal{N}_d(\xi, \sigma^2 I)$ by conditioning on $||Y|| = 1$. The parameters of the distributions are related as

$$\mu = \xi/||\xi||, \quad \kappa = ||\xi||/\sigma^2.$$

The TM algorithm can thus be applied by giving Y_i a Gaussian distribution as above and conditioning on $||Y_i|| = 1$ for all i. This yields

$$\xi_0 = s/N \text{ and } \sigma_0^2 = \frac{\sum ||y_i||^2 - (\sum y_i)^2/N}{Nd} = (1 - r^2/N^2)/d.$$

Expressing this in terms of the parameters of the von Mises–Fisher distribution, and using that $E(Y \mid ||Y|| = 1) = \mu A_d(\kappa)$, yields the TM algorithm as

$$\mu_n = s/r, \qquad \kappa_n = dr\rho_n/\{N^2 + \rho_n(\rho_n - 2)r^2\},$$

where $\rho_0 = 1$ and $\rho_{n+1} = \rho_n + 1 - NA_d(\kappa_n)/r$.

5 Discussion

Although the TM algorithm was developed with the chain graph models in mind, its structure promises potential application in more general contexts. For any given family of models used for studying the effects of explanatory variables on responses, one needs to augment the parameter space by equipping the explanatory variables with a suitable marginal distribution in such a way that the joint distribution becomes simple. Admittedly, this is often more difficult than the data augmentation needed for the EM algorithm.

We emphasise that, when the augmented joint family can be chosen in different ways, the analysis in § 2 indicates that one should choose this family to be as rich as possible, to prevent strong 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, as in the last model in Table 3, and the algorithm will converge after a single cycle.

The TM algorithm is also potentially applicable in cases where a conditional maximum likelihood estimate is preferred to the full maximum likelihood estimate, for example to eliminate undesirable effects of nuisance parameters.

Note that, as the entire action of the algorithm happens in the mean value space, as captured in (5), 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 TM and EM algorithms can be used when values of the response variables are missing. Iteration (5) should then be replaced by

$$\tau_{n+1} = \tau_n + E_{\theta_n}\{U \mid x, y_{\mathrm{obs}}\} - E_{\theta_n}\{U \mid x\},$$

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

Acknowledgement

This research was supported in part by the ESPRIT Programme. We are grateful to Rolf Sundberg, Mike Titterington and two anonymous referees for constructive comments concerning both exposition and substance, including the suggestion of a suitable name for the algorithm.

References

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

Cox, D. R. & Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall.

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

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

Edwards, D. (1990). Hierarchical interaction models (with Discussion). *J. R. Statist. Soc.* B **52**, 3–20, 51–72.

Edwards, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. New York: Springer-Verlag.

Frydenberg, M. & Edwards, D. (1989). A modified iterative proportional scaling algorithm for estimation in regular exponential families. *Comp. Stat. Data Anal.* **8**, 143–53.

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

Geng, Z., Wan, K. & Tao, F. (2000). Mixed graphical models with missing data and the partial imputation EM algorithm. *Scand. J. Statist.* **27**, 433–44.

Jebara, T. & Pentland, A. (1998). Maximum conditional likelihood via bound maximisation and the CEM algorithm. In *Advances in Neural Information Processing Systems 11*, Ed. M. S. Kearns, S. A. Solla & D. A. Cohn. Cambridge, MA: MIT Press.

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

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

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

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.

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

Mardia, K. V. (1975). Distribution theory for the von Mises–Fisher distribution and its application. In *Statistical Distributions for Scientific Work*, Ed. G. P. Patil, S. Kotz & J. K. Ord), pp. 113–30. Boston: Reidel.

Marschner, I. C. (2001). On stochastic versions of the EM algorithm. *Biometrika* **88**, 281–286.

Meng, X.-L. & van Dyk, D. (1997). The EM algorithm — an old folk song sung to a fast new tune (with Discussion). *J. R. Statist. Soc.* B **59**, 511–67.

Ortega, J. M. & Rheinboldt, W. C. (1970). *Iterative Solutions of Nonlinear Equations in Several Variables.* New York: Academic Press.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vettering, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Wermuth, N. & Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with Discussion). *J. R. Statist. Soc.* B **52**, 21–72.

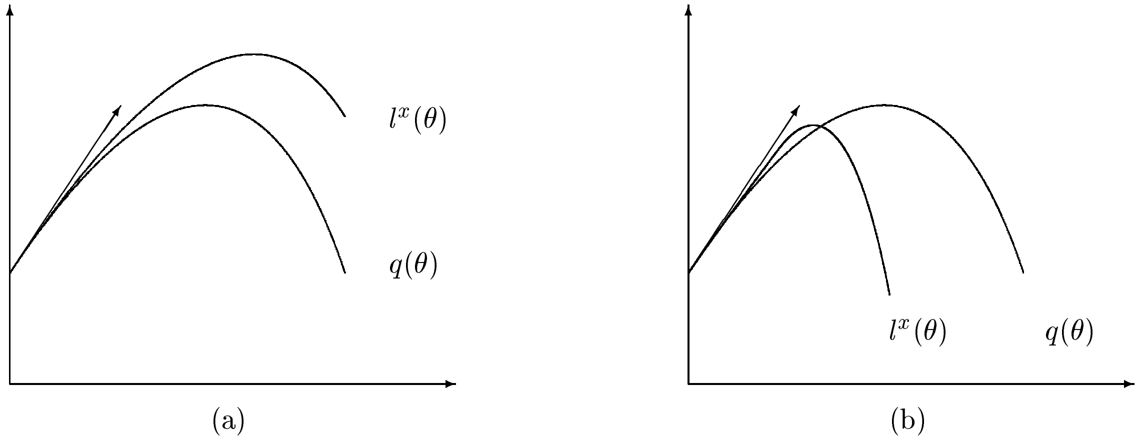


Figure 1: The relationship between a translate of the conditional likelihood function l^x and its approximation q obtained by linearising the marginal likelihood function. In (a), l^x lies entirely above q and any value which increases q will also increase l^x. In (b), corresponding to strong negative observed marginal information, maximisation of q will decrease l^x.

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

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

I	J	(y^k, z^k)						
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 3: Progress of the TM-algorithm with different CG-regression models.

Model 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	$I, J/IY, JY, IZ, JZ/IY, JZ$	I, Y	4	no
8	$IJ/IJY, IJZ/Y, Z$	I, J	1	no