The convergence rate of the TM algorithm of Edwards and Lauritzen

by

Rolf Sundberg,

Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden

rolfs@matematik.su.se

10 January, 2002

SUMMARY

Edwards & Lauritzen (2001) have recently proposed the TM algorithm for finding the maximum likelihood estimate when the likelihood can be truly or artificially regarded as a conditional likelihood, and the full likelihood is more easily maximised. They have presented a proof of convergence, provided that the algorithm is supplemented by a line search. In this note a simple expression, in terms of observed information matrices, is given for the convergence rate of the algorithm per se, when it converges, and the result elucidates also in which situations the algorithm will require a line search. Essentially these are cases when the full model does not adequately fit the data.

Some key words: Conditional likelihood; Exponential family; Graphical chain model; Iterative method; Maximum likelihood estimation.

1 Introduction

Maximum likelihood estimation in mixed graphical chain models usually requires an iterative method for the solution of the likelihood equations. With such situations in mind, Edwards & Lauritzen (2001) have recently proposed a general method, called the TM method, for iterative likelihood maximisation when the likelihood can be regarded as a conditional likelihood within a full, joint model that allows an explicit solution. They have found the algorithm to be a useful tool in complex CG-regression models, which are building blocks in graphical chain models, and they suggest that the algorithm might be of use also in other situations. However, they have not proved that the algorithm will always converge even if the starting point is close to the maximum likelihood estimate, unless the algorithm is supplied with a line search, and they have not investigated the convergence rate theoretically. Here we will supplement their study with a simple result concerning this rate. From the same result it will be seen that the algorithm can be expected to converge if the full model constructed is an adequate description of data, but not necessarily otherwise. The result is general, with no explicit reference to graphical chain models. A simple toy example, admitting an explicit solution, and a logistic regression will be used for illustration.

2 Rates of convergence

2.1 Application of Ostrowski's theorem.

We will use essentially the same notation as in Edwards & Lauritzen (2001), and likewise we implicitly assume suitable regularity conditions. Some response data y are given, whose distribution depends on some x, and we introduce a distribution for x, so as to be able to write the loglikelihood to be maximised as a conditional loglikelihood l^x, in terms of a full loglikelihood l and a marginal loglikelihood l_x:

$$l^x(\theta; y|x) = l(\theta; x, y) - l_x(\theta; x).$$

The TM algorithm for maximising l^x starts from some θ_0 that is successively updated by the following two-step procedure.

T-step. Given θ_n, calculate the tilted loglikelihood $q_n(\theta) = l(\theta) - \theta^{\mathrm{T}} \dot{l}_x(\theta_n)$, where $\dot{l}$ with or without sub/superscripts is used like l to denote the corresponding Fisher score functions, as in Edwards & Lauritzen (2001).

M-step. Maximise $q_n(\theta)$ to find θ_{n+1}.

If this algorithm converges, it will be to a root $\hat{\theta}$ of the original likelihood equations, since the derivative $\dot{q}_n$ of q_n at $\theta_{n+1} = \theta_n = \theta$ satisfies

$$\dot{q}_n(\theta) = \dot{l}(\theta; x, y) - \dot{l}_x(\theta; x) = \dot{l}^x(\theta; y|x),$$

which is zero precisely when $\theta = \hat{\theta}$. The algorithm may be expressed in the form $\theta_{n+1} = g(\theta_n)$, with the vector-valued composite function $g = \dot{l}^{-1} \circ \dot{l}_x$, and $\hat{\theta}$ is a fixed point of g, i.e. $\hat{\theta} = g(\hat{\theta})$. The speed of convergence of the algorithm close to $\hat{\theta}$ depends on the function g as described in the so-called Ostrowski's theorem (Ostrowski, 1960, Ch. 18).

CRITERION (Ostrowski). *For a fixed point $\hat{\theta}$ to be a point of attraction of the algorithm $\theta_{n+1} = g(\theta_n)$, a sufficient condition is that the Jacobian matrix of g at the point $\hat{\theta}$ has all its eigenvalues numerically less than 1, and a necessary condition is that they are numerically at most 1. The geometric rate of convergence is the numerically largest eigenvalue of this Jacobian.*

The geometric interpretation is that the eigenvalues identify the rate of convergence in the eigenvector directions, and the error in the direction of slowest convergence will dominate after some iterations. To apply this criterion, note that the Jacobian matrix for the TM algorithm is $J^{-1}J_x = I - J^{-1}J^x$, where J denotes the observed information matrix and I the identity matrix, and the matrices are calculated in terms of the maximum likelihood estimate $\hat{\theta}$. The eigenvalues of this Jacobian are all real. This follows from the fact that J and J^x can be simultaneously diagonalised, as soon as J^x is positive definite at $\hat{\theta}$, and otherwise in combination with the discussion in §2.3 below.

Provided J^x is positive definite at $\hat{\theta}$, it further follows that not only J^x but also J will be positive definite at $\hat{\theta}$, and as a consequence that all eigenvalues of $J^{-1}J_x = I - J^{-1}J^x$ at $\hat{\theta}$ are less than $+1$. The positivity of J is seen as follows. In each M-step of the algorithm, the function $q_n(\theta)$ is maximised, and at the maximum it will have a negative definite Hessian matrix. Since q_n differs by only a linear term from $l(\theta)$, the full model observed information J will be positive definite, not only at $\hat{\theta}$ but in fact at each successive θ_n on the way to $\hat{\theta}$.

There are some analogies with the EM algorithm, for which the corresponding Jacobian matrix is found in Sundberg (1976) and in Theorem 4 of Dempster et al. (1977), and is seen to depend on marginal and conditional information matrices.

We have seen above that the rate of convergence of the TM algorithm is the numerically largest eigenvalue of $J^{-1}J_x$ at $\hat{\theta}$, provided this value is less than 1. Under regularity conditions, including the requirement that data contain information about the whole of θ in the conditional model, we have also seen that all eigenvalues are less than $+1$. In many applications there will be eigenvalues which are $+1$, however, corresponding to parameters in the full model which are absent in the conditional model. This case requires a more detailed discussion, found in §2.3 below. Before that, we investigate in §2.2 the theoretically simpler case when the full and conditional models have the same parameters, in order to consider more in detail when the convergence criterion will be satisfied.

2.2 Case 1. The full model does not introduce additional parameters.

The argumentation in the previous section showed that, only provided data contain information about the whole of θ in the conditional model, all eigenvalues of $J^{-1}J_x = I - J^{-1}J^x$ at $\hat{\theta}$ are strictly less than $+1$. If the observed information matrices were replaced by their expected values at the same parameter point, namely the Fisher information matrices, we could even conclude that all eigenvalues would be nonnegative, and convergence of the algorithm would follow. It would be nice if we could argue that the observed information matrices were similar to the expected versions, as is usual with large datasets. However, in the present situation the marginal model need not have any connection with the reality behind data, and then this need not at all be true. Hence, the following condition can be crucial: the data (x, y) should reasonably fit not only the conditional, but also the full model.

Note that the model of interest is the conditional model, and the full model can be, and often will be, an artificial construct, just for the sake of making the algorithm applicable. If data do not fit the full model there is no guarantee that $J_x(\hat{\theta})$ will be positive semidefinite, or even that the eigenvalues of $J^{-1}J_x$ will exceed -1. Not even if data are in effect generated from the full model is there any guarantee that the algorithm will converge. This is illustrated in the following toy example, which has only two parameters and allows an explicit solution.

Example 1. Proportional regression. Let y_i be independent and distributed as $N(\lambda x_i, \sigma^2)$, for fixed x_i. In order to apply the algorithm, artificially imagine the x_i to be sampled from a $N(0, \sigma^2)$ distribution, with the same σ. Elementary calculations show that, with $\theta = (\lambda, \sigma^2)$,

$$J(\hat{\theta}) = n \begin{pmatrix} S_{xx}/\sigma^2 & 0 \\ 0 & S_{xx}/\sigma^6 \end{pmatrix}, \qquad J^x(\hat{\theta}) = n \begin{pmatrix} S_{xx}/\sigma^2 & 0 \\ 0 & 0.5/\sigma^4 \end{pmatrix},$$

where $S_{xx} = \sum x_i^2/n$. The marginal J_x is the difference $J - J^x$, so its last element is the only nonzero element, $n(S_{xx} - 0.5\sigma^2)/\sigma^6$. Hence, the eigenvalues of $J^{-1}J_x$ are 0 and $1 - 0.5\hat{\sigma}^2/S_{xx}$. The zero eigenvalue reflects the fact that the algorithm will find $\hat{\lambda}$ already in the first iteration. If the full model is reasonable we should expect that $S_{xx} \doteq \hat{\sigma}^2$, and consequently the other eigenvalue to be about $1 - 0.5 = 0.5$. This means that the deviation from $\hat{\sigma}^2$ will be halved for each iteration if we are close to $\hat{\sigma}^2$. However, if the full model is inadequate for describing data, in such a way that $S_{xx} << \hat{\sigma}^2$, the nonzero eigenvalue may be less than -1, and the algorithm will then diverge from $\hat{\sigma}^2$, even if we start close to this value. A line search along the line from θ_n to θ_{n+1}, as devised by Edwards & Lauritzen (2001), would be successful, however, since the error at θ_n is alternating.

Apart from the fact that no iterative algorithm is really needed in this example, an intuitively more natural choice of a full model might seem to be one in which $\text{var}(x_i)$ is allowed be a new, free parameter, τ^2 say. This brings us into Case 2, which is the topic of the next section.

2.3 Case 2. The full model introduces additional parameters.

In most applications of interest the full model will involve more parameters than the conditional model. Let $\theta = (\alpha, \beta)$, where α is the parameter of the conditional model, so that

$$l(\theta; x, y) = l^x(\alpha; y|x) + l_x(\alpha, \beta; x).$$

The algorithm can be expressed in the form

$$\dot{l}(\theta_{n+1}) = \dot{l}(\theta_n) - \dot{l}^x(\theta_n).$$

Here $\dot{l}^x$ is the gradient with respect to the full θ, but, since l^x does not depend on β, the β-component of $\dot{l}^x$ is zero. This means that the β-component $\dot{l}_\beta$ of $\dot{l}$ remains constant under the algorithm. In other words, the algorithm moves on a surface $\dot{l}_\beta =$ constant. If the algorithm is started from a point for which $\dot{l}_\beta = 0$, e.g. from the maximum likelihood estimate of the full model, then the constant will be zero. Thus, β_n will be the profile maximum likelihood estimate in the full model when $\alpha = \alpha_n$. This restricted maximisation will be seen in $J^{-1}J_x$ as a unit eigenvalue of multiplicity $\dim(\beta)$. The interpretation is that the convergence rate of the algorithm is the numerically largest eigenvalue of $J^{-1}J_x$ after removal of an eigenvalue 1 of multiplicity $\dim(\beta)$. The corresponding eigenvectors span the β-space with zero coefficients for α. This is seen by noting that the $\dim(\beta)$ last columns of $J^{-1}J^x$ are zero. The other eigenvalues of $J^{-1}J^x$ are in effect the same as the eigenvalues of the α-corner of $J^{-1}J^x$. Hence, if we let subscript $(.)_{\alpha\alpha}$ denote the α-corner of a matrix, the convergence rate can be expressed as the numerically largest eigenvalue of

$$I_{\alpha\alpha} - (J^{-1})_{\alpha\alpha}(J^x)_{\alpha\alpha}. \tag{1}$$

In a situation when the statistic x induces a cut, completely separating the parameters of the conditional and the marginal likelihoods, $(J_x)_{\alpha\alpha}$ will be identically zero, or equivalently $(J^x)_{\alpha\alpha} = (J)_{\alpha\alpha}$, and the eigenvalues of (1) will all be exactly zero. In connection with graphical chain models, a cut would signal a desirable simplification. From another perspective, however, if the full model has been constructed just because we did not want to solve the conditional model likelihood equations directly, a cut would be unfortunate, because it would lead us back to the original equations in the M-step. As an example, if, in Example 1, $\text{var}(x_i)$ had been allowed to be a new, free parameter, the statistic x would have induced a cut, and the M-step would have demanded us to solve the original likelihood equation.

Example 2. Logistic regression. Edwards & Lauritzen (2001) illustrate the convergence of the TM algorithm in a simple logistic regression of y on x, for some data with clearly controlled x-values. An artificial distribution for x is obtained by pretending that, for each given outcome 0 or 1 of the y-variable, the x-variables are normally distributed, with different mean values but the same variance. This means that the marginal distribution for x is taken to be a mixture of the two normal distributions. This increases the parameter dimension from 2 to 4. The TM algorithm is seen in their Table 1 to converge at a rate of about $-1/4$; that is the error is alternating and reduced by a factor of 1/4 at each iteration. Numerical calculation of $J^{-1}J_x$ at the point $\hat{\theta}$ showed that the eigenvalues for the two-dimensional logistic parameter were -0.26 and $+0.06$. At the starting point θ_0 the corresponding values were -0.29 and $+0.09$, so the starting point gave a good indication of the convergence rate at $\hat{\theta}$. There were also the two eigenvalues of $+1$ for the two additional parameters of the full model, but as explained above they are of no relevance for the desired convergence. The negative eigenvalue can be taken as reflecting the fact that the normal distribution for x given y was not a reasonable model.

2.4 Aspects of line search

As we have seen above, the matrix $J^{-1}J_x$ at $\hat{\theta}$ can have eigenvalue(s) less than -1, and then the TM algorithm will not converge. Edwards & Lauritzen (2001) show for this case that a line search along the line connecting θ_n and θ_{n+1} ensures convergence. Let then the algorithm modified with line search be

$$\theta_{n+1} = (1-\lambda)\,\theta_n + \lambda\, g(\theta_n) \tag{2}$$

for a fixed λ, $0 < \lambda \leq 1$, where g is the same function as before, defining the pure TM algorithm. It is seen that the new Jacobian matrix is a λ-weighted average of the previous one and the identity matrix. Hence the eigenvalues controlling the rate of convergence are moved towards the value $+1$, for λ small enough all eigenvalues will exceed -1, and the line search will make the algorithm converge, if it starts close enough to $\hat{\theta}$. The locally optimal λ-value in (2) will depend on the two outermost eigenvalues of $J^{-1}J_x$ at $\hat{\theta}$, η_{min} and η_{max} say, and is

$$\lambda_{opt} = 1/\{1 - (\eta_{min} + \eta_{max})/2\},$$

with a corresponding rate of convergence $\lambda_{opt}(\eta_{max} - \eta_{min})/2$. Note how the line search benefits from the fact that the algorithm is alternating if it is divergent, that is if $\eta_{min} \leq -1$. Note also that there is not much need to change a satisfactory λ-value from one iteration to the next, at least not in a vicinity of $\hat{\theta}$, whereas Edwards & Lauritzen have in mind a λ varying with n.

3 Explicit formulae for exponential families

It is instructive to see how the rate of convergence can be expressed when the conditional distribution of Y given X is an exponential family which is embedded in a wider exponential family for (X, Y). Note that the corresponding marginal distribution of X need not be of exponential type. We use the canonical parametrisation, and similarly to Edwards & Lauritzen (2001) we write the canonical statistic for the full model as $T = (U, V)$, where $U = u(X, Y)$ and $V = v(X)$. Here U is the canonical statistic in the conditional model. As an example consider logistic regression, as in Example 2 above. The norming constant of the logistic conditional model for each observation y, given x, is $C_x(\alpha_1, \alpha_2) = 1 + \exp(\alpha_1 + \alpha_2 x)$, where α_1 and α_2 are the regression parameters. If we choose as marginal distribution for x for example a mixture of two exponentials or, as Edwards & Lauritzen (2001) do, a mixture of two Gaussian distributions with common variance, a three- or four-parametric joint distribution for (x, y) is created where the conditional norming constant C_x fits into the exponential part of the full model.

Standard theory for exponential families implies that $\dot{l} = T - E_\theta(T)$, and $\dot{l^x} = U - E_\theta(U|X)$ augmented by $\dim(\beta)$ zero components. It follows that the TM algorithm can be written in the form

$$\begin{aligned} E_{\theta_{n+1}}(U) &= E_{\theta_n}(U) + u - E_{\theta_n}(U|x), \qquad (3) \\ E_{\theta_{n+1}}(V) &= E_{\theta_n}(V) = v. \end{aligned}$$

If this is expressed in the full-model mean value parametrisation for U, $\tau = E_\theta(U)$, this looks even simpler, and is given in this form in Edwards & Lauritzen (2001, §3): $\tau_{n+1} = \tau_n + u - E_{\theta_n}(U|x)$.

Further standard exponential family theory gives that, for any θ, $J = \mathrm{var}(T)$ and the α-corner of J^x equals $\mathrm{var}(U|x)$. From well-known expressions for inverted matrices, or alternatively by transformation to a mixed parametrisation, with α and $E(V)$ as orthogonal parameters, it follows that the rate-determining part of the matrix $J^{-1}J_x$, as given more generally in (1), can be expressed as

$$I_{\alpha\alpha} - \{\mathrm{var}(U) - \mathrm{cov}(U,V)\mathrm{var}(V)^{-1}\mathrm{cov}(V,U)\}^{-1}\mathrm{var}(U|x), \qquad (4)$$

where $I_{\alpha\alpha}$ is the unit matrix of dimension $\dim(\alpha)$. Here we may first conclude that the eigenvalues of (4) are necessarily upper-bounded by $+1$, so if the algorithm per se does not converge, a line search from a good starting point will be successful. The inverse in front of $\mathrm{var}(U|x)$ approximates the inverse of the conditional variance $\mathrm{var}(U|V)$, if a normal approximation of (U,V) is reasonable. Since $\mathrm{var}(U|V) = E\{\mathrm{var}(U|X)|V\} + \mathrm{var}\{E(U|X)|V\}$ we have some reason to expect the eigenvalues of (4) to be nonnegative. However, this argument of course does not even guarantee that all eigenvalues exceed -1, in particular not when the full model yields an inadequate description of the data.

Remark. If we cannot observe y completely but only have some incomplete data z, which can be regarded as a function $z = f(y)$, then we may simply substitute $E_{\theta_n}(U|z,x)$ for u in (3), as suggested by Edwards & Lauritzen (2001), to form a hybrid of the TM and EM algorithms. Since the derivative of $E(U|z,x)$ is $\mathrm{var}(U|z,x)$

(Sundberg, 1974), the rate-determining matrix of the hybrid algorithm is changed from (4) for the pure TM algorithm to

$$I_{\alpha\alpha} - \{\mathrm{var}(U) - \mathrm{cov}(U, V)\mathrm{var}(V)^{-1}\mathrm{cov}(V, U)\}^{-1}\{\mathrm{var}(U|x) - \mathrm{var}(U|z, x)\}.$$

The last factor, $\mathrm{var}(U|x) - \mathrm{var}(U|z, x)$, is necessarily nonnegative at the maximum point, being the observed information in the conditional model. Note that, if the eigenvalues of (4) are all nonnegative, they will be closer to 1 under incomplete data, so the algorithm will be slower.

4 Concluding discussion

We have seen in this paper how the rate of convergence of the TM algorithm can be calculated as the numerically largest eigenvalue of a matrix specified. It is not so important to be able to calculate this rate of convergence when the data are already available and the algorithm can be tried without the effort of calculating the information matrices. Also, in principle, we should know the point $\hat{\theta}$ where the rate should be calculated, so we would have to run the algorithm and would then see its rate of convergence empirically. However, calculating the rate at the starting point might give a good indication of the rate at the point $\hat{\theta}$, as in Example 2 above.

A more important role for the rate of convergence results concerns the possibility of comparing different devices. There may be several different candidates for the full model, with more or less restricted parametrisations, and they could be compared theoretically. The rate of convergence will differ with x, in particular depending on how well the full model is able to fit the given (x, y), as in Example 1 above. In the example of their §4.2, Edwards & Lauritzen (2001) touch on the choice between two augmented models inducing the same CG-regression model for y given x; only one of them was found to require line search. For such studies the results can be useful, perhaps by explicit theoretical calculations in very simple cases, but more likely by extensive computer investigations.

Finally the above results contribute to our general understanding of when the TM algorithm will converge, and when it will not. In particular, we saw that it can be crucial that the constructed full model fits the data reasonably.

References

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.

Edwards, D. & Lauritzen, S.L. (2001). The TM algorithm for maximising a conditional likelihood function. *Biometrika* **88**, 961-72.

Ostrowski, A.M. (1960). *Solution of Equations and Systems of Equations.* New York: Academic Press.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.* **1**, 49-58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Commun. Statist.* B **5**, 55-64.