Statistical Methods in Medical Research 1997; **6**: 38–54

EM algorithms without missing data

Mark P Becker Department of Biostatistics, University of Michigan, Ann Arbor, Michigan
Ilsoon Yang Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts and **Kenneth Lange** Departments of Biostatistics and Mathematics, University of Michigan, Ann Arbor, Michigan, USA

Most problems in computational statistics involve optimization of an objective function such as a loglikelihood, a sum of squares, or a log posterior function. The EM algorithm is one of the most effective algorithms for maximization because it iteratively transfers maximization from a complex function to a simple, surrogate function. This theoretical perspective clarifies the operation of the EM algorithm and suggests novel generalizations. Besides simplifying maximization, optimization transfer usually leads to highly stable algorithms with well-understood local and global convergence properties. Although convergence can be excruciatingly slow, various devices exist for accelerating it. Beginning with the EM algorithm, we review in this paper several optimization transfer algorithms of substantial utility in medical statistics.

1 Introduction

Medical statistics employs a broad array of models for description, analysis and inference. In estimating parameters, most of these models require optimization of an objective function such as a loglikelihood, a sum of squares, a penalized loglikelihood or a log posterior function. Some loglikelihoods are relatively simple to optimize, for example those encountered in generalized linear models with canonical link functions. Other loglikelihoods are inherently more nonlinear and consequently more challenging. Although Newton's method and its statistical cousin Fisher scoring are routinely used to maximize well-behaved loglikelihoods, both algorithms have their drawbacks. Newton's method entails calculation of complicated second derivatives and is equally happy to head toward a minimum or saddlepoint as it is toward a maximum. Scoring requires calculation of the expected information matrix. Outside exponential families of distributions, this task is often impossible. For problems with large numbers of parameters, both algorithms involve large matrix inversions. It is hardly surprising that statisticians find the simplicity and numerical stability of the EM algorithm appealing. The EM algorithm is based on an optimization transfer principle that replaces a complex optimization problem by a sequence of simpler ones. In this paper we argue that optimization transfer rather than missing data is the key ingredient of the EM algorithm. We illustrate this point of view by presenting several algorithms that involve no missing data, but otherwise mimic the general behaviour of the EM algorithm.

In discussing the EM algorithm one should keep in mind that it is not so much an algorithm as a prescription for constructing an algorithm. In Section 2, we review the theoretical underpinnings of the EM algorithm and illustrate its application to latent class models for the analysis of diagnostic accuracy. In many cases either the E-step or

Address for correspondence: Mark P Becker, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA. Email: mbecker@umich.edu

 0962-2802(97)SM125RA

the M-step of the EM algorithm is intractable. The EM gradient algorithm provides a straightforward remedy for the failure of the M step. We discuss the EM gradient algorithm in conjunction with an application to mixed logistic regression. Stochastic sampling provides a computationally more expensive remedy for the failure of the E-step.[1] These two devices and variations on the EM theme such as Bayesian EM,[2] the ECM algorithm[3] and accelerated EM[4–6] (this is also discussed in an unpublished manuscript by M Jamshidian and RL Jennrich) all belong in the armoury of every research statistician.

More esoteric but equally worth knowing is the central role of convexity in the EM algorithm. The ascent property of the EM algorithm ultimately depends on convexity through the entropy inequality. We hope to convince readers that even in the absence of missing data convexity can be exploited to create optimization transfer algorithms. The analogies between the EM algorithm and other optimization transfer algorithms are so strong that a unified theory can be erected to cover both local and global convergence. In Section 3 we discuss the desirable properties that optimization transfer often shares with the EM algorithm. Section 4 describes several concrete examples. Naturally, design of these new algorithms relies on art as much as science. However, the same can be said for the EM algorithm when one reflects on the clever missing data structures that stand behind many specific EM algorithms.

2 The EM algorithm

2.1 Overview

At the heart of the classical EM algorithm[7] is the notion of missing or incomplete data, which can consist of missing observations in the ordinary sense or theoretically missing entities concocted by the statistician specifically to simplify optimization. Let Y denote the *observed data*, Z the *missing data*, and $X = (Y, Z)$ the *complete data*. The EM algorithm, like all maximum likelihood algorithms, seeks to maximize the log-likelihood $L(\theta)$ of the observed data with respect to a vector of unknown parameters θ. If $f(X|\theta)$ denotes the density function of the complete data, then the EM algorithm maximizes the surrogate function

$$Q(\theta|\theta^n) = E[\ln f(X|\theta)|Y, \theta^n]$$

with respect to its left argument. This gives the update θ^{n+1} of the current iterate θ^n in the search for the maximum likelihood estimate $\hat{\theta}$. The essence of good optimization transfer is that maximizing $Q(\theta|\theta^n)$ is much simpler than maximizing $L(\theta)$. If this is not the case, then the statistician has chosen the wrong missing data structure. The price of simplification by optimization transfer is iteration. Formation of the conditional expectation of the complete data loglikelihood $\ln f(X|\theta)$ given the observed data Y and the current parameter vector θ^n constitutes the E-step of the EM algorithm. Maximization of this conditional expectation $Q(\theta|\theta^n)$ constitutes the M-step. A surprising feature of the EM optimization transfer is that increasing $Q(\theta|\theta^n)$ forces an increase in $L(\theta)$. This ascent property holds because $L(\theta) - Q(\theta|\theta^n)$ attains its minimum at $\theta = \theta^n$. In view of this fact, we can argue that

$$\begin{aligned} L(\theta^{n+1}) &= Q(\theta^{n+1}|\theta^n) + [L(\theta^{n+1}) - Q(\theta^{n+1}|\theta^n)] \\ &\geq Q(\theta^n|\theta^n) + [L(\theta^n) - Q(\theta^n|\theta^n)] \\ &= L(\theta^n) \end{aligned}$$

with strict inequality when $\theta^{n+1} \neq \theta^n$.

Proof that $L(\theta) - Q(\theta|\theta^n)$ attains its minimum at $\theta = \theta_n$ hinges on the entropy inequality

$$E_g[\ln g] \geq E_g[\ln h] \tag{2.1}$$

which is a direct consequence of Jensen's inequality.[8,9] In inequality (2.1) g and h denote probability densities with respect to a measure μ, and E_g denotes expectation with respect to the probability measure $gd\mu$. Equality occurs if and only if $g = h$ except for a set of measure zero. If we assume that $l(y|\theta) = e^{L(\theta)}$ is the density of the observed data Y and apply (2.1) with g equal to the conditional density $f(x|\theta^n)/l(y|\theta^n)$ and h equal to the conditional density $f(x|\theta)/l(y|\theta)$, then we find that

$$\begin{aligned} Q(\theta|\theta^n) - L(\theta) &= E(\ln \left[\frac{f(X|\theta)}{l(Y|\theta)}\right] | Y = y, \theta^n) \\ &\leq E(\ln \left[\frac{f(X|\theta^n)}{l(Y|\theta^n)}\right] | Y = y, \theta^n) \\ &= Q(\theta^n|\theta^n) - L(\theta^n) \end{aligned}$$

2.2 Application to latent class models for diagnostic accuracy

Sensitivity and specificity are two measures routinely used to assess the accuracy of diagnostic tests or diagnosticians in medical research. Sensitivity is the probability of a 'positive' test result given that the patient has the disease, while specificity is the probability of a 'negative' test result given that the patient does not have the disease. Both sensitivity and specificity can be calculated directly when there exists a definitive reference test. Latent class analysis has been applied to assess diagnostic accuracy when it is impossible to calculate estimates of sensitivity and specificity directly.[10–16] The premise of the latent class model is that the tests are imperfect indicators of the unobserved true disease status, which is treated as a latent variable. Responses within a latent class are assumed to be independent. Departures from independence in the observed table of test outcomes occur as the result of mixing the two unobserved latent tables. Here we consider the situation in which inference for four diagnostic tests are of interest. (Unfortunately, three or fewer tests renders the following model unidentifiable.[17]) Let the diagnostic tests A, B, C and D each have two categories, 'positive' (1) or 'negative' (0), indexed by i, j, k and l, respectively. The cell frequencies in the crossclassification of the results of the tests follow a multinomial distribution with 2^4 cells. Let y_{ijkl} and μ_{ijkl} denote the observed and expected cell frequencies, respectively, and let p_{ijkl} and π_{ijkl} denote the observed and expected cell probabilities, respectively. In this notation we have $y_{ijkl} = Np_{ijkl}$ and $\mu_{ijkl} = N\pi_{ijkl}$, where $N = \sum_{ijkl} y_{ijkl} = \sum_{ijkl} \mu_{ijkl}$ is the sample size. Finally, let T denote the latent variable for the true disease status, i.e. $T = 1$ if a patient has the disease, and $T = 0$ otherwise. If T is indexed by t, then π_{ijkl} can be decomposed as

$$\pi_{ijkl} = \sum_{t=0}^{1} \pi_{ijklt}^{ABCDT} = \sum_{t=0}^{1} \pi_t^T \pi_{ijkl|t}^{ABCD|T} \tag{2.2}$$

where π_t^T is the probability that a patient has disease status t and $\pi_{ijkl|t}^{ABCD|T}$ is the conditional probability that he or she shows test results (i,j,k,l) given disease status t. Clearly

$$\sum_{t=0}^{1} \pi_t^T = 1, \qquad \sum_{ijkl} \pi_{ijkl|t}^{ABCD|T} = 1, \qquad t = 0,1$$

The sensitivity and specificity of test A are

$$\pi_{1|1}^{A|T} = \pi_{1+++|1}^{ABCD|T} \quad \text{and} \quad \pi_{0|0}^{A|T} = \pi_{0+++|0}^{ABCD|T}$$

respectively, where a plus sign denotes summation over a corresponding subscript. The sensitivities and specificities for the other tests are defined similarly. The EM algorithm is well suited for maximum likelihood estimation with latent class models. Indeed, Goodman[17] developed the specific EM algorithm now described several years before Dempster *et al*. enunciated the general EM algorithm in 1977.[7] The complete data x_{ijklt} include a hidden indicator t of disease status for each patient represented in the observed data y_{ijkl}. If we make the natural local independence assumption that the response variables (A,B,C,D) are independent conditional on the disease status t of a patient, then

$$\pi_{ijkl|t}^{ABCD|T} = \pi_{i|t}^{A|T} \pi_{j|t}^{B|T} \pi_{k|t}^{C|T} \pi_{l|t}^{D|T}, \qquad i,j,k,l,t = 0, 1 \tag{2.3}$$

This translates into the complete data loglikelihood

$$\sum_{ijklt} x_{ijklt} \ln (\pi_{ijklt}^{ABCDT}) = \sum_{ijklt} x_{ijklt} \ln (\pi_t^T \pi_{i|t}^{A|T} \pi_{j|t}^{B|T} \pi_{k|t}^{C|T} \pi_{l|t}^{D|T})$$

The EM algorithm permits straightforward estimation of the parameters π_t^T, $\pi_{i|t}^{A|T}$, $\pi_{j|t}^{B|T}$, $\pi_{k|t}^{C|T}$, and $\pi_{l|t}^{D|T}$ of the model as summarized in:

- E-step. The expected values of the complete data are imputed as

$$x_{ijklt}^n = y_{ijkl} \frac{\pi_{ijklt}^{ABCDTn}}{\pi_{ijkl}^{ABCDn}}$$

  conditional on the observed data and the current estimated cell probabilities π^n.
- M-step. The surrogate function $Q(\pi|\mathbf{y}, \pi^n)$

$$Q(\pi|\mathbf{y}, \pi^n) = \sum_{ijklt} x_{ijklt}^n \ln (\pi_t^T \pi_{i|t}^{A|T} \pi_{j|t}^{B|T} \pi_{k|t}^{C|T} \pi_{l|t}^{D|T})$$

  is maximized with respect to the parameters, yielding for example

$$\pi_t^{Tn} = \frac{x_{++++t}^n}{N}$$
$$\pi_{i|t}^{A|Tn} = \frac{x_{i+++t}^n}{x_{++++t}^n}$$

Exact solution of the M-step is possible in this example because the surrogate function separates the various parameters. Thus, the EM algorithm transforms a complex, nonlinear optimization problem with equality and boundary constraints into a sequence of simple optimization problems with exact solutions. This is an extremely attractive feature of the EM algorithm when it occurs, and it does for a large number of interesting problems.[7,18]

2.3 EM gradient algorithm

The EM gradient algorithm[19] is ideally suited to problems where the M-step of the EM algorithm cannot be solved exactly. A natural candidate for solving the M-step in such cases is Newton's method. Because Newton's method converges quickly (at a quadratic rate) while the EM algorithm converges slowly (at a linear rate), there is little point in taking multiple Newton's steps within each M-step. Thus, the EM gradient algorithm iterates according to

$$\begin{aligned}\theta^{n+1} &= \theta^n - d^{20}Q(\theta^n|\theta^n)^{-1}\, d^{10}Q(\theta^n|\theta^n) \\ &= \theta^n - d^{20}Q(\theta^n|\theta^n)^{-1}\, dL(\theta^n)\end{aligned}$$

where the operator d^{ij} takes the ith partial derivative with respect to the left argument and the jth partial derivative with respect to the right argument of Q. Substitution of $dL(\theta^n)$ for $d^{10}Q(\theta^n|\theta^n)$ is valid because $L(\theta) - Q(\theta|\theta^n)$ attains its minimum at $\theta = \theta^n$. At the optimal point $\hat{\theta}$, the EM gradient algorithm map shares with the EM algorithm map the differential

$$I - d^{20}Q(\hat{\theta}|\hat{\theta})^{-1}\, d^2L(\hat{\theta}) = -d^{20}Q(\hat{\theta}|\hat{\theta})^{-1}[d^2L(\hat{\theta}) - d^{20}Q(\hat{\theta}|\hat{\theta})] \tag{2.4}$$

Since the dominant eigenvalue of the differential of an algorithm map determines the local rate of convergence of the algorithm in a neighbourhood of $\hat{\theta}$, the EM and EM gradient algorithms behave almost identically. One can even show that the EM gradient algorithm obeys the ascent condition $L(\theta^{n+1}) \geq L(\theta^n)$ near $\hat{\theta}$.[19]

2.4 Application to mixed logistic regression

Follmann and Lambert[20] employ mixed logistic regression to analyse the dose–response experiments of Ashford and Walker[21] on trypanosomes – protozoans causing sleeping sickness and nagana. Here we consider a simple version of their more general nonparametric models. To accommodate the overdispersion in the trypanosome data, Follmann and Lambert postulate an underlying dichotomous latent variable T with two classes $t = 1, 2$ having probabilities π_1^T and $\pi_2^T = 1 - \pi_1^T$. The observed data are frequencies y_{ij}, where $i = 1, \ldots, s$ indexes the dose levels and $j = 1, 2$ indexes the response (dead or alive). The complete data are frequencies x_{ijt} conveying latent class status as well as dose level and response. If we let π_{ij}^Y denote the probabilities of the

observable responses and $\pi_{ij|t}^{Y|T}$ denote the conditional probability of response j given latent class t at dose level i, then we can write the loglikelihood of the observed data as

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{s}\sum_{j=1}^{2} y_{ij} \ln(\pi_{ij}^{Y}) \\ &= \sum_{i=1}^{s}\sum_{j=1}^{2} y_{ij} \ln\left(\sum_{t=1}^{2} \pi_t^T \pi_{ij|t}^{Y|T}\right) \end{aligned}$$

For the complete data we assume a linear logistic regression model of the form

$$\ln \frac{\pi_{i1|t}^{Y|T}}{\pi_{i2|t}^{Y|T}} = \beta_{0t} + \beta_1 D_i$$

the covariate D_i being $\ln(dose_i)$. The model has parameter vector $\theta = (\beta_{01}, \beta_{02}, \beta_1, \pi_1^T)$, and complete data loglikelihood

$$\sum_{t=1}^{2} x_{++t} \ln(\pi_t^T) + \sum_{t=1}^{2}\sum_{i=1}^{s}\sum_{j=1}^{2} x_{ijt} \ln(\pi_{ij|t}^{Y|T}) \tag{2.5}$$

The E-step of the EM algorithm replaces the x_{ijt} by their expected values

$$x_{ijt}^n = y_{ij} \frac{\pi_t^{Tn} \pi_{ij|t}^{Y|Tn}}{\pi_{ij}^{Yn}}$$

conditional on the observed data and the current parameter vector θ^n. The M-step of the EM algorithm immediately yields the update

$$\pi_t^{T,n+1} = \frac{x_{++t}^n}{N}$$

where N is the number of subjects. Estimation of the logistic regression parameters can be accomplished by iteratively weighted least squares using the imputed data x_{ijt}^n.[22] However, iterating within each M step defeats the simplicity of the EM algorithm. The EM gradient algorithm now comes to the rescue and suggests that we apply one step of Newton's method to that part of $Q(\theta|\theta^n)$ captured by the triple sum in (2.5) with the imputed data x_{ijt}^n replacing the complete data x_{ijt}. We omit the mechanics of Newton's method because for the most part they appear in Section 4.1. To summarize, we recommend in this problem using the exact solution for $\pi_t^{T,n+1}$ and applying one step of Newton's method to update the remaining parameters. Such a hybrid algorithm is consistent with the venerable dictum of numerical analysis that one should approximate only when absolutely necessary.

3 Desirable features of optimization transfer

Despite its shortcomings, Newton's method is the gold standard for optimization algorithms in computational statistics. Besides leading to a fast, quadratic rate of convergence in a neighbourhood of the global maximum, Newton's method automatically provides the asymptotic covariance matrix of the parameter estimates. The price exacted for these advantages include its failure to distinguish local maxima, local minima, and saddlepoints from the global maximum when it is started too far from the global maximum, the necessity of computing second derivatives, and the chore of inverting the observed information matrix. Inversion of the observed information matrix is particularly problematic if it is ill-conditioned or exceptionally large. For example, in emission tomography the number of parameters is typically on the order of 10^4.[2,23–25] Optimization transfer as we construe it mimics the EM algorithm by constructing a surrogate function $Q(\theta|\theta^n)$ and maximizing it with respect to its left argument. This action gives the updated iterate θ^{n+1} to the current iterate θ^n in maximizing an objective function $L(\theta)$. We drop from the EM paradigm the requirement of viewing $Q(\theta \mid \theta^n)$ as a conditional expectation, but we retain the requirement that the difference $L(\theta) - Q(\theta \mid \theta^n)$ achieve its minimum at $\theta = \theta^n$. The latter condition is the key ingredient in proving the ascent property $L(\theta^{n+1}) \geq L(\theta^n)$ that lends numerical stability to the algorithm. In Section 4 we illustrate how appropriate surrogate functions can be constructed by exploiting convexity features of $L(\theta)$.

Each of the optimization transfer algorithms discussed in the sequel enjoys good global and local convergence properties.[7,19,26,27] Each converges to the global maximum if the objective function $L(\theta)$ is strictly concave. If the objective function is not concave, but all stationary points are isolated, then an optimization transfer algorithm is guaranteed to converge to one of the stationary points. This stationary point need not be a local maximum. In unusual circumstances, even the ordinary EM algorithm will converge to a saddlepoint.[26,27] If we cannot maximize $Q(\theta \mid \theta^n)$ exactly, then the EM gradient algorithm is available. To ensure that the EM gradient algorithm works properly, we require that the Hessian matrix $d^{20}Q(\theta^n|\theta^n)$ be negative definite and that a limited line search be conducted in the Newton direction from the current point θ^n. With these provisos, the EM gradient algorithm also converges to one of the stationary points, provided these are isolated. Finally, the local rate of convergence of an optimization transfer algorithm (in either its exact or EM gradient forms) is determined by the dominant eigenvalue of the differential (2.4).[7,19]

As noted in the Introduction, optimization transfer often substitutes a simple optimization problem for a difficult one. In doing so it can achieve one or more of the following objectives: (i) avoid large matrix inversions; (ii) linearize the optimization problem; (iii) separate the parameters of the optimization problem; and (iv) handle equality and inequality constraints gracefully. All of these advantages are nicely illustrated in the examples considered in Section 4. These examples include: (a) Böhning and Lindsay's quadratic lower bound principle,[28,29] (b) Dutter and Huber's optimization transfer for elliptically symmetric distributions,[30] (c) an adaptive barrier method for convex programming,[31] (d) application of De Pierro's first convexity argument to image reconstruction in transmission tomography[5,23,32] and (e) extension

of De Pierro's second convexity argument to optimization transfer for generalized linear models with canonical link functions, probit regression, multinomial regression, and least L_1 regression.[23] To our knowledge, example (e) is new.

4 Examples of optimization transfer algorithms

4.1 Quadratic lower bound principle

Böhning and Lindsay[29] introduced a lower bound algorithm under the assumption that a negative definite matrix B can be found such that $d^2L(\theta) - B$ is nonnegative definite for all θ. They set

$$Q(\phi|\theta) = L(\theta) + dL(\theta)^t(\phi - \theta) + \frac{1}{2}(\phi - \theta)^t B(\phi - \theta)$$

where the superscript t indicates a transpose operation. Since

$$L(\phi) = L(\theta) + dL(\theta)^t(\phi - \theta) + \frac{1}{2}(\phi - \theta)^t d^2L(\bar{\theta})(\phi - \theta)$$

for some intermediate point $\bar{\theta}$ between ϕ and θ, it follows that

$$\begin{aligned} L(\phi) - Q(\phi|\theta) &= \frac{1}{2}(\phi - \theta)^t[d^2L(\bar{\theta}) - B](\phi - \theta) \\ &\geq 0 \end{aligned}$$

Clearly $L(\phi) - Q(\phi|\theta)$ attains its minimum at $\phi = \theta$. The quadratic lower bound algorithm amounts to maximizing $L(\theta)$ by Newton's method with B substituted for $d^2L(\theta)$. Böhning and Lindsay[28,29] apply the quadratic lower bound principle to logistic regression, multinomial logistic regression, mixture models, and Cox's proportional hazards model. Here we illustrate the implementation of the algorithm on logistic regression. Let z_i denote a (large) vector of predictors for each observation y_i, and let $\langle z_i, \theta \rangle = \sum_j z_{ij}\theta_j$; $i = 1, \ldots, m$. The y_i are assumed to be realizations of independent Bernoulli random variables with success probabilities

$$\pi_i = \frac{\exp(\langle z_i, \theta \rangle)}{1 + \exp(\langle z_i, \theta \rangle)}$$

The loglikelihood, score, and the observed information are

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{m} [y_i \ln \pi_i + (1 - y_i)\ln(1 - \pi_i)] \\ dL(\theta) &= \sum_{i=1}^{m} (y_i - \pi_i) z_i \\ -d^2L(\theta) &= \sum_{i=1}^{m} \pi_i(1 - \pi_i) z_i z_i^t \end{aligned}$$

Because $\pi_i(1 - \pi_i) \leq 1/4$ for each i, the nonpositive definite matrix $B = -\sum_{i=1}^{m} \frac{1}{4} z_i z_i^t$ is designed so that $d^2L(\theta) - B$ is nonnegative definite.

4.2 Dutter and Huber algorithm

Dutter and Huber[30] introduced an optimization transfer algorithm for elliptically symmetric densities

$$\frac{e^{-1/2}\kappa(\delta^2)}{c_\kappa \det(\Omega)^{1/2}}$$

on R^k, where c_κ is a normalizing constant, $\delta^2 = (y-\mu)^t\Omega^{-1}(y-\mu)$, and $\kappa(s)$ is a increasing, strictly concave function. The multivariate t is a well-known example of an elliptically symmetric distribution[33,34] If $y_1,\ldots,y_m$ is a sequence of independent realizations from the density (3.8) with location vectors $\mu_1,\ldots,\mu_m$ and scale matrices $\Omega_1,\ldots,\Omega_m$, then the surrogate for the actual loglikelihood $L(\theta)$ is the normal loglikelihood

$$Q(\theta|\theta^n) = -\frac{1}{2}\sum_{i=1}^{m}\{w_i(\theta^n)\delta_i^2(\theta) + \ln\det[\Omega_i(\theta)]\} \tag{4.1}$$

where $w_i(\theta^n) = \kappa'[\delta_i^2(\theta^n)]$ is a weight associated with the ith observation. Note that the difference $L(\theta) - Q(\theta|\theta^n)$ attains its minimum at θ^n because $\kappa'(s^n)s - \kappa(s)$ attains its minimum at $s = s^n$. The array of techniques from linear algebra and multivariate analysis for maximizing the normal loglikelihood can be brought to bear on maximizing $Q(\theta|\theta^n)$.

As a simple illustration of the Dutter and Huber algorithm, consider least L_p regression.[35] If the independent realizations $y_1,\ldots,y_m$ have unit variances and $0 < p \le 2$, then the choice $\kappa(s) = s^{p/2}$ leads to least L_p regression. The Dutter and Huber algorithm minimizes at each iteration

$$\sum_{i=1}^{m} w_i(\theta^n)[y_i - \mu_i(\theta)]^2$$

with weights $w_i(\theta^n) = |y_i - \mu_i(\theta^n)|^{p-2}$. In other words, least L_p regression can be done by iteratively reweighted least squares. A problem with this algorithm is that infinite weights occur for those observations with zero residuals. Redefining the weights as

$$w_i(\theta^n) = \frac{1}{\epsilon + |y_i - \mu_i(\theta^n)|^{2-p}}$$

for a small $\epsilon > 0$ overcomes this difficulty. This choice of weights corresponds to

$$\kappa'(s) = \frac{1}{\epsilon + s^{1-p/2}}$$

and also leads to a maximum likelihood algorithm. For $p = 1$ the slightly revised algorithm minimizes the criterion

$$\sum_{i=1}^{m}[|\, y_i - \mu_i(\theta)\,| - \epsilon \ln(\epsilon + |\, y_i - \mu(\theta)\,|)]$$

4.3 Transmission tomography

In transmission tomography, high energy photons are sent from an external source through the body to an external detector. In statistical image reconstruction, the plane region of an X-ray slice is divided into small rectangular pixels, and pixel j is assigned an attenuation coefficient θ_j. Each photon sent from the source along projection i (line of flight) has probability $\exp(-\langle l_i, \theta \rangle)$ of avoiding absorption by the body, where l_i is the vector of intersection lengths l_{ij} of the ith projection with the jth pixel. For a Poisson number (mean d_i) of photons departing along projection i, the number y_i of photons detected is Poisson with mean $d_i \exp(-\langle l_i, \theta \rangle)$. Since different projections are independent, the loglikelihood reduces to

$$L(\theta) = \sum_{i=1}^{m} [-d_i \exp(-\langle l_i, \theta \rangle) - y_i \langle l_i, \theta \rangle] \tag{4.2}$$

Note the nonnegativity constraints $\theta_j \geq 0$ and $l_{ij} \geq 0$. The loglikelihood in (4.2) can be abbreviated as $L(\theta) = \sum_i f_i(\langle l_i, \theta \rangle)$ using the strictly concave functions $f_i(s) = -d_i e^{-s} - y_i s$. Following De Pierro's lead[32] in emission tomography, define the admixture constants

$$\lambda_{ij} = \frac{l_{ij} \theta_j^n}{\langle l_i, \theta^n \rangle}$$

Since $\sum_j \lambda_{ij} = 1$ and $f_i(s)$ is concave

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{m} f_i \left(\sum_j \lambda_{ij} \frac{\theta_j}{\theta_j^n} \langle l_i, \theta^n \rangle \right) \\ &\geq \sum_i \sum_j \lambda_{ij} f_i \left(\frac{\theta_j}{\theta_j^n} \langle l_i, \theta^n \rangle \right) \\ &= Q(\theta | \theta^n) \end{aligned}$$

with equality when $\theta_j = \theta_j^n$ for all j. Thus, the difference $L(\theta) - Q(\theta|\theta^n)$ attains its minimum of 0 when $\theta = \theta^n$. By construction, maximization of $Q(\theta|\theta^n)$ separates into a sequence of one-dimensional problems, each of which can be solved approximately by one step of Newton's method.[5]

4.4 Linear and convex programming

The standard convex programming problem is to minimize $f(\theta)$ subject to a set of linear constraints $A\theta = b$ and nonnegativity constraints $\theta \geq 0$. Interior point methods seek the minimum while remaining on the interior $\{\theta : A\theta = b,\ \theta > 0\}$ of the feasible region. Minimization can be transferred to the surrogate function

$$Q(\theta|\theta^n) = f(\theta) - \mu \sum_i \left[\theta_i^n \ln \theta_i - \theta_i \right] \tag{4.3}$$

for $\mu > 0$. The adaptive barrier term $\mu \sum_i [\theta_i^n \ln \theta_i - \theta_i]$ on the right of (4.3) has its

maximum at $\theta = \theta^n$ and forces θ^{n+1} to have all components positive. Of course, no component is prevented from tending to 0 as n tends to ∞. Lange[31] and Iusem and Teboulle[36] independently proposed this optimization transfer algorithm, which applies regardless of whether $f(\theta)$ is convex. A single step of Newton's method can be used to approximately minimize $Q(\theta|\theta^n)$ subject to $A\theta = b$ and $\theta > 0$. The update in this case is

$$\theta^{n+1} = \theta^n - G^n(I - A^t[AG^nA^t]^{-1}AG^n)\, df(\theta^n)$$
$$G^n = [d^2f(\theta^n) + \mu D^n]^{-1}$$

where D^n is the diagonal matrix with ith diagonal element $1/\theta_i^n$. As an example consider the linear programming problem of Klee and Minty,[37] which requires minimizing θ subject to the inequality constraints $0 \leq \theta_1 \leq 1$ and $\beta\theta_{i-1} \leq \theta_i \leq 1 - \beta\theta_{i-1}$ for $i = 2, \ldots, m$. This problem illustrates the exponential complexity of the simplex algorithm. Started at the point $\theta = (0.001, \ldots, 0.001)^t$ when $m = 8$ and $\beta = 1/4$, the new algorithm achieves the minimum at $(0, \ldots, 0, 1)^t$ to four significant digits in 11 iterations and to seven significant digits in 18 iterations.

4.5 De Pierro's second convexity argument

To accommodate a smoothing penalty in emission tomography reconstructions, De Pierro[23] introduced a second method for optimization transfer. In contrast to the multiplicative technique discussed in our transmission tomography example, his second technique is additive. Here we extend it to certain generalized linear models,[22] multinomial regression, and least L_1 regression. Our point of departure is the maximization of a sum of the form

$$L(\theta) = \sum_{i=1}^{m} f_i(\langle z_i, \theta \rangle) \tag{4.4}$$

where the functions $f_i(r)$ are strictly concave in the real variable r, z_i is a vector of k covariates for the ith of m observations, and θ is a parameter vector of length k. In generalized linear modelling, $L(\theta)$ represents the loglikelihood of m independent observations from a regular exponential family. In least L_1 regression, $L(\theta)$ is the negative sum of m absolute residuals $|y_i - \langle z_i, \theta \rangle|$. We consider first the smooth functions $f_i(r)$ encountered in generalized linear models and multinomial regression. After digesting this case, we turn to the nondifferentiable functions $f_i(r)$ of least L_1 regression and derive an algorithm distinct from that in Section 4.2.

If the $f_i(r)$ are twice continuously differentiable, then the first and second differentials of $L(\theta)$ are

$$dL(\theta) = \sum_{i=1}^{m} f_i'(\langle z_i, \theta \rangle) z_i$$
$$d^2L(\theta) = \sum_{i=1}^{m} f_i''(\langle z_i, \theta \rangle) z_i z_i^t$$

Provided each $f''(r)$ is strictly negative, a necessary and sufficient condition that $L(\theta)$

be strictly concave is that column vectors z_i form a covariate matrix z of full rank k. If $L(\theta)$ is strictly concave and a stationary point $\hat{\theta}$ exists, then $\hat{\theta}$ furnishes the global maximum of $L(\theta)$.[38]

To effect an optimization transfer, we could use the quadratic lower bound principle with the matrix B defined by

$$B = \sum_{i=1}^{m} \inf_r f_i''(r) z_i z_i^t$$

In examples such as Poisson regression, this procedure fails because $\inf_r f_i''(r) = -\infty$ for each i. Even when the quadratic lower bound principle succeeds, inversion of the matrix B poses a problem when the number of parameters is large. Alternatively, we can exploit convexity and choose nonnegative numbers λ_{ij} such that $\sum_j \lambda_{ij} = 1$ and $\lambda_{ij} > 0$ whenever $z_{ij} \neq 0$. Possible candidates for λ_{ij} are $\lambda_{ij} = z_{ij}^2/||z_i||_2^2$, $\lambda_{ij} = |z_{ij}|/||z_i||_1$ and $\lambda_{ij} = 1/|U_i|$ for $j \in U_i$ and 0 for $j \notin U_i$, where $U_i = \{j : z_{ij} \neq 0\}$. In view of the concavity of the f_i, if we let $S_i = \{j : \lambda_{ij} > 0\}$, then

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{m} f_i(\langle z_i, \theta \rangle) \\ &= \sum_{i=1}^{m} f_i\left[\sum_{j \in S_i} \lambda_{ij} \frac{z_{ij}}{\lambda_{ij}} (\theta_j - \theta_j^n) + \langle z_i, \theta^n \rangle\right] \\ &\geq \sum_{i=1}^{m} \sum_{j \in S_i} \lambda_{ij} f_i\left[\frac{z_{ij}}{\lambda_{ij}} (\theta_j - \theta_j^n) + \langle z_i, \theta^n \rangle\right] \qquad (4.5) \\ &= Q(\theta|\theta^n) \end{aligned}$$

with equality when $\theta = \theta^n$. In the surrogate function $Q(\theta|\theta^n)$, all parameters are separated. In most cases it is impossible to carry out these one-dimensional maximizations explicitly. If we resort to the EM gradient algorithm, then we update θ_j^n by

$$\theta_j^{n+1} = \theta_j^n - \left[\sum_{i \in T_j} f_i''(\langle z_i, \theta^n \rangle) \frac{z_{ij}^2}{\lambda_{ij}}\right]^{-1} \sum_{i \in T_j} f_i'(\langle z_i, \theta^n \rangle) z_{ij} \qquad (4.6)$$

where $T_j = \{i : \lambda_{ij} > 0\}$ and the f_i are assumed twice continuously differentiable. As a first application, consider generalized linear models with canonical link functions. If r denotes the canonical parameter, then the canonical link assumption takes $f_i(r) = y_i r - a(r)$ for each observation $Y_i = y_i$ from the underlying exponential family. In this setting the mean and variance of Y_i are $\mu_i(r) = a_i'(r)$ and $v_i(r) = a''(r)$,[22] and equation (4.6) reduces to

$$\theta_j^{n+1} = \theta_j^n + \left[\sum_{i \in T_j} v_i(\langle z_i, \theta^n \rangle) \frac{z_{ij}^2}{\lambda_{ij}}\right]^{-1} \sum_{i \in T_j} [y_i - \mu_i(\langle z_i, \theta^n \rangle)] z_{ij}$$

Examples are:

(A) *Normal distribution* (identity link)

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i \in T_j}(y_i - \langle z_i, \theta^n \rangle) z_{ij}}{\sum_{i \in T_j} z_{ij}^2 / \lambda_{ij}}$$

Estimation of the variance σ^2 separates from estimation of θ. Conventionally, statisticians use the unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m (y_i - \langle z_i, \hat{\theta} \rangle)^2}{m - k}$$

rather than the maximum likelihood estimator.

(B) *Binomial distribution* (logit link)

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i \in T_j}(y_i - N_i \pi_i^n) z_{ij}}{\sum_{i \in T_j} N_i \pi_i^n (1 - \pi_i^n) z_{ij}^2 / \lambda_{ij}}$$

Here $\pi_i^n = e^{\langle z_i, \theta^n \rangle} / (1 + e^{\langle z_i, \theta^n \rangle})$ is the success probability per trial, and N_i is the number of trials for the ith case.

(C) *Poisson distribution* (log link)

$$\theta_j^{n+1} = \theta_j^n + \frac{\sum_{i \in T_j}(y_i - E_i e^{\langle z_i, \theta^n \rangle}) z_{ij}}{\sum_{i \in T_j} E_i e^{\langle z_i, \theta^n \rangle} z_{ij}^2 / \lambda_{ij}}$$

Here $\mu_i = E_i e^{\langle z_i, \theta \rangle}$ is the mean of the ith case, E_i being a fixed and known offset.

The algorithm (4.6) is also applicable to generalized linear models with non-canonical link functions provided the $f_i(r)$ are strictly concave. A case in point is the probit model for Bernoulli variation. If $\Phi(r)$ denotes the standard normal distribution function, then the relevant functions

$$f_i(r) = y_i \ln \Phi(r) + (1 - y_i) \ln [1 - \Phi(r)]$$

are known to be strictly concave.[39] Another example is the gamma model with mean $\alpha e^{\langle z_i, \theta \rangle}$, where α is the fixed shape parameter. Here we have $f_i(r) = -\alpha r - y_i e^{-r}$. This parameterization is more convenient than the canonical parameterization because it guarantees positivity of the mean.

Multinomial regression models are not generalized linear models, but they do belong to the more general family of exponential dispersion models.[40] The loglikelihood for a loglinear model with count y_i in the ith of m cells is

$$L(\theta) = \sum_{i=1}^m y_i \langle z_i, \theta \rangle - N \ln \left[\sum_{i=1}^m e^{\langle z_i, \theta \rangle} \right]$$

where $N = \sum_{i=1}^m y_i$. The inequality

$$\begin{aligned}\sum_{i=1}^{m} \exp\left(\langle z_i, \theta\rangle\right) &\leq \sum_{j=1}^{k} \sum_{i \in T_j} \lambda_{ij} \exp\left[\frac{z_{ij}}{\lambda_{ij}}(\theta_j - \theta_j^n) + \langle z_i, \theta^n\rangle\right] \\ &= \sum_{j=1}^{k} \sum_{i \in T_j} \lambda_{ij} g_{ij}(\theta_j | \theta^n) \\ &= \sum_{j=1}^{k} g_j(\theta_j | \theta^n)\end{aligned}$$

based on the convexity of e^r implies that

$$\begin{aligned}L(\theta) &\geq \sum_{i=1}^{m} y_i \langle z_i, \theta\rangle - N \ln\left[\sum_{j=1}^{k} g_j(\theta_j | \theta^n)\right] \\ &= Q(\theta | \theta^n)\end{aligned}$$

with equality at $\theta = \theta^n$. Although this surrogate maximization function $Q(\theta|\theta^n)$ does not separate parameters, it does yield simple one-step Newton updates. The first differential $d^{10}Q(\theta|\theta^n)$ of $Q(\theta|\theta^n)$ has entries

$$\frac{\partial Q(\theta|\theta^n)}{\partial \theta_j} = \sum_{i=1}^{m} y_i z_{ij} - N \sum_{i \in T_j} \frac{g_{ij}(\theta_j|\theta^n) z_{ij}}{\sum_{l=1}^{k} g_l(\theta_l|\theta^n)}$$

and the second differential $-d^{20}Q(\theta|\theta^n)$ is a nonnegative definite matrix that can be expressed as a rank-one perturbation of a diagonal matrix. Computation of the inverse of $-d^{20}Q(\theta|\theta^n)$ is therefore straightforward using the Sherman–Morrison formula.[41]

In least L^1 regression, the functions $f_i(r) = -|y_i - r|$ are concave but not differentiable. Minimization of the surrogate function in (4.4) with separated parameters reduces to solving for each j the minimization problem

$$\min_{\theta_j} \sum_{i \in T_j} w_i |d_i - \theta_j|$$

where $w_i = |z_{ij}|$ and

$$d_i = \theta_j^n + (y_i - \langle z_i, \theta^n\rangle)\frac{\lambda_{ij}}{z_{ij}}$$

Statisticians will immediately recognize the solution as the median of the discrete random variable taking the value d_i with probability proportional to the weight w_i.

In all of the examples discussed in this section, the optimization transfer algorithm avoids matrix inversion. This is a major advantage in problems with many parameters. The primary shortcoming of the algorithm is that it can exhibit the same painfully slow convergence seen in the EM algorithm. Our limited experience to date suggests that acceleration techniques based on conjugate gradients and quasi-Newton methods help a great deal.[42] Further work on acceleration of these algorithms and on the optimal selection of the λ_{ij} is certainly warranted.

5 Discussion

Theoretical development and practical application of the EM algorithm have emphasized the statistical concept of missing data. This notion can reflect missing observations in the ordinary sense or theoretically hidden random variables. The E-step of the algorithm fills in the missing data and constructs a surrogate function $Q(\theta|\theta^n)$ for the loglikelihood $L(\theta)$. The M-step maximizes $Q(\theta|\theta^n)$ with respect to its left argument to give the next iterate θ^{n+1}. Statisticians have exercised great creativity in identifying appropriate missing data structures. The resulting algorithms often give intuitively appealing parameter updates that incorporate parameter constraints gracefully. Many statisticians, the current authors included, have been seduced by the technical challenges of constructing EM algorithms.

While we do not want to deprecate these creative outlets, we have argued here that the strength of the EM algorithm lies not so much in its exploitation of missing data structures as in its optimization transfer interpretation. The ascent property and the convergence behaviour of the EM algorithm depend on optimization transfer, not on missing data. In constructing a surrogate function $Q(\theta|\theta^n)$ for a loglikelihood or more general objective function $L(\theta)$, the key requirement is that the difference $L(\theta) - Q(\theta|\theta^n)$ achieves its minimum at $\theta = \theta^n$. The examples covered in Section 4 and, indeed, the classical EM algorithm itself illustrate the crucial role of convexity in defining appropriate surrogate functions. Almost all of the well-known inequalities in mathematics revolve around convexity as well. It is our belief that statisticians will eventually derive as much pleasure and profit from defining optimization transfer algorithms based on convexity as they have from identifying missing data structures.

If simplicity and elegance are the hallmarks of the EM algorithm, then an often painfully slow rate of convergence is its Achilles heel. We would be remiss if we failed to mention at least a few of the devices for accelerating the EM algorithm. Early on, Louis[6] suggested Aitken acceleration. This proposal has had more theoretical than practical impact. More recently, Jamshidian and Jennrich[4] advocated generalized conjugate gradients. This device reduces iteration counts by one or two orders of magnitude in many hard problems. Similar and even more spectacular accelerations can be achieved by combining quasi-Newton techniques[43] with the EM gradient algorithm.[44] Such hybrid algorithms are particularly attractive because they retain the robust behaviour of the underlying optimization transfer algorithm during early iterations while taking advantage of the rapid quadratic rate of convergence of Newton's method during late iterations. Quasi-Newton accelerations also accommodate parameter constraints more naturally than generalized conjugate gradients. Nonetheless, it is premature to declare victory in the battle to improve the performance of the EM algorithm. This is still an area in need of more research. Good algorithm design, here as elsewhere in biostatistics, is as relevant as ever.

Acknowledgements

The authors gratefully acknowledge support from the National Institutes of Health (Grants CA53787 and GM53275) and the Statistics Center at Cornell University.

References

1 Sobel E, Lange K. Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* 1993; **2**: 263–82.

2 Green P. Bayesian reconstruction for emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging* 1990; **9**: 84–94.

3 Meng X-L, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 1993; **80**: 267–78.

4 Jamshidian M, Jennrich RI. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association* 1993; **88**: 221–28.

5 Lange K, Fessler JA. Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Transactions on Image Processing* 1995; **4**: 1430–38.

6 Louis TA. Finding the observed information using the EM algorithm. *Journal of the Royal Statistical Society Series B* 1982; **44**: 98–130.

7 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 1977; **39**: 1–38.

8 Billingsley P. *Probability and measure*, 3rd edition. New York: John Wiley, 1995.

9 Brown LD. *Fundamentals of statistical exponential families with applications in statistical decision theory*, IMA Lecture Notes – Monograph Series, Volume 9, Hayward, CA: Institute of Mathematical Statistics, 1986.

10 Clogg CC. Some latent structure models for the analysis of Likert-type data. *Social Science Research* 1979; **8**: 287–301.

11 Clogg CC. Latent class models. In: Arminger G, Clogg CC, Sobel ME, eds. *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum, 1995: 311–59.

12 Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* 1979; **28**: 20–28.

13 Espeland MA, Handelman SA. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 1989; **45**: 587–99.

14 Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5**: 21–27.

15 Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. *Statistics in Medicine* 1990; **9**: 559–72.

16 Young MA. Evaluating diagnostic criteria: a latent class paradigm. *Journal of Psychiatric Research* 1983; **17**: 285–96.

17 Goodman, LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**: 215–31.

18 McLachlan GJ, Krishnan T. *The EM algorithm and extensions* New York: John Wiley, 1997.

19 Lange K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society Series B* 1995; **57**: 425–37.

20 Follman D, Lambert D. Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* 1989; **84**: 295–300.

21 Ashford R, Walker PJ. Quantal response analysis for a mixture of populations. *Biometrics* 1972; **28**: 981–88.

22 McCullagh P, Nelder JR. *Generalized linear models*, 2nd edition. London: Chapman & Hall, 1989.

23 De Pierro AR. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Transactions on Medical Imaging* 1995; **14**: 132–37.

24 Levitan E, Herman G. A maximum *a posteriori* probability expectation maximization algorithm for image reconstruction in emission tomography. *IEEE Transactions on Medical Imaging* 1987; **6**: 185–92.

25 Shepp L, Vardi Y. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* 1982; **1**: 113–21.

26 Boyles RA. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society Series B* 1983; **45**: 47–50.

27 Wu CF. On the convergence properties of the EM algorithm. *Annals of Statistics* 1983; **11**: 95–103.

28 Böhning D. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics* 1992; **44**: 197–200.

29 Böhning D, Lindsay BG. Monotonicity of quadratic approximation algorithms. *Annals of the Institute of Statistical Mathematics* 1988; **40**: 641–63.

30 Dutter R, Huber PJ. Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation* 1981; **13**: 79–113.

31 Lange, K. An adaptive barrier method for convex programming. *Methods and Applications of Analysis* 1994; **4**: 392–402.
32 De Pierro AR. On the relation between the ISRA and EM algorithm for positron emission tomography. *IEEE Transactions on Medical Imaging* 1993; **12**: 328–33.
33 Dempster AP, Laird NM, Rubin DB. Iteratively reweighted least squares for linear regression when the errors are normal/independent distributed. In: Krishnaiah PR ed. *Multivariate analysis – V*. Amsterdam: North Holland, 1980.
34 Lange K, Little RJA, Taylor JMG. Robust statistical modeling using the *t* distribution. *Journal of the American Statistical Association* 1989; **84**: 881–96.
35 Lange K, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics* 1993; **2**: 175–98.
36 Iusem AN, Teboulle M. Convergence rate analysis of nonquadratic proximal methods for convex and linear programming. *Mathematics of Operations Research* 1995; **20**: 657–77.
37 Klee V, Minty GJ. How good is the simplex algorithm? In *Inequalities, III* (Proceedings of the Third Symposium, University of California, Los Angeles, California, 1969; dedicated to the memory of Theodore S Motzkin). New York: Academic Press, 1972: 159–75.
38 Dennis JE, Jr, Schnabel RB. *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs, NJ: Prentice–Hall, 1983.
39 Silvapulle, MJ. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society Series B* 1981; **43**: 310–13.
40 Joergensen B. Exponential dispersion models, with discussion. *Journal of the Royal Statistical Society Series B* 1987; **49**: 127–45.
41 Miller KS. *Some eclectic matrix theory*. Malabar, FL: Robert E Krieger, 1987.
42 Yang I. Latent class marginal models for the analysis of cross-classified categorical data. PhD dissertation, University of Michigan, 1996.
43 Gill PE, Murray W, Wright MH. *Practical optimization*. San Diego, CA: Academic Press, 1981.
44 Lange K. A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica* 1995; **5**: 1–18.