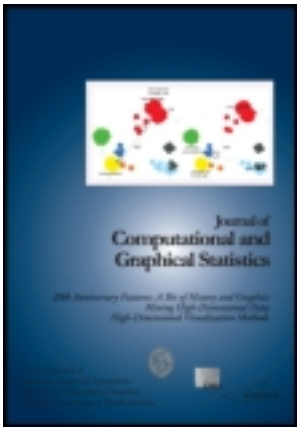

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

MM Algorithms for Some Discrete Multivariate Distributions

Hua Zhou[a] & Kenneth Lange[a]

[a] Hua Zhou is Post-Doctoral Fellow, Department of Human Genetics, University of California, Los Angeles, CA 90095-7088 . Kenneth Lange is Professor, Departments of Biomathematics, Human Genetics, and Statistics, University of California, Los Angeles, CA 90095-7088.

Published online: 01 Jan 2012.

To cite this article: Hua Zhou & Kenneth Lange (2010) MM Algorithms for Some Discrete Multivariate Distributions, Journal of Computational and Graphical Statistics, 19:3, 645-665, DOI: 10.1198/jcgs.2010.09014

To link to this article: http://dx.doi.org/10.1198/jcgs.2010.09014

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

MM Algorithms for Some Discrete Multivariate Distributions

Hua ZHOU and Kenneth LANGE

The MM (minorization–maximization) principle is a versatile tool for constructing optimization algorithms. Every EM algorithm is an MM algorithm but not vice versa. This article derives MM algorithms for maximum likelihood estimation with discrete multivariate distributions such as the Dirichlet-multinomial and Connor–Mosimann distributions, the Neerchal–Morel distribution, the negative-multinomial distribution, certain distributions on partitions, and zero-truncated and zero-inflated distributions. These MM algorithms increase the likelihood at each iteration and reliably converge to the maximum from well-chosen initial values. Because they involve no matrix inversion, the algorithms are especially pertinent to high-dimensional problems. To illustrate the performance of the MM algorithms, we compare them to Newton's method on data used to classify handwritten digits.

Key Words: Dirichlet and multinomial distributions; Inequalities; Maximum likelihood; Minorization.

1. INTRODUCTION

The MM algorithm generalizes the celebrated EM algorithm (Dempster, Laird, and Rubin 1977). In this article we apply the MM (minorization–maximization) principle to devise new algorithms for maximum likelihood estimation with several discrete multivariate distributions. A series of research papers and review articles (Groenen 1993; de Leeuw 1994; Heiser 1995; Hunter and Lange 2004; Lange 2004; Wu and Lange 2010) have argued that the MM principle can lead to simpler derivations of known EM algorithms. More importantly, the MM principle also generates many new algorithms of considerable utility. Some statisticians encountering the MM principle for the first time react against its abstraction, unfamiliarity, and dependence on the mathematical theory of inequalities. This is unfortunate because real progress can be made applying a few basic ideas in a unified framework. The current article relies on just three well-known inequalities. For most of our

Hua Zhou is Post-Doctoral Fellow, Department of Human Genetics, University of California, Los Angeles, CA 90095-7088 (E-mail: *huazhou@ucla.edu*). Kenneth Lange is Professor, Departments of Biomathematics, Human Genetics, and Statistics, University of California, Los Angeles, CA 90095-7088.

Journal of Computational and Graphical Statistics, Volume 19, Number 3, Pages 645–665
DOI: 10.1198/jcgs.2010.09014

examples, the derivation of a corresponding EM algorithm appears much harder, the main hindrance being the difficulty of choosing an appropriate missing data structure.

Discrete multivariate distributions are seeing wider use throughout statistics. Modern data mining employs such distributions in image reconstruction, pattern recognition, document clustering, movie rating, network analysis, and random graphs. High-dimension data demand high-dimensional models with ten to hundreds of thousands of parameters. Newton's method and Fisher scoring are capable of finding the maximum likelihood estimates of these distributions via the parameter updates

$$\theta^{(n+1)} = \theta^{(n)} + M\big(\theta^{(n)}\big)^{-1} \nabla L\big(\theta^{(n)}\big),$$

where $\nabla L(\theta)$ is the score function and $M(\theta)$ is the observed or the expected information matrix, respectively. Several complications can compromise the performance of these traditional algorithms: (a) the information matrix $M(\theta)$ may be expensive to compute, (b) it may fail to be positive definite in Newton's method, (c) in high dimensions it is expensive to solve the linear system $M(\theta)x = \nabla L(\theta^{(n)})$, and (d) if parameter constraints and parameter bounds intrude, then the update itself requires modification. Although mathematical scientists have devised numerous remedies and safeguards, these all come at a cost of greater implementation complexity. The MM principle offers a versatile weapon for attacking optimization problems of this sort. Although MM algorithms have at best a linear rate of convergence, their updates are often very simple. This can tip the computational balance in their favor. In addition, MM algorithms are typically easy to code, numerically stable, and amenable to acceleration. For the discrete distributions considered here, there is one further simplification often missed in the literature. These distributions involve gamma functions. To avoid the complications of evaluating the gamma function and its derivatives, we fall back on a device suggested by Haldane (1941) that replaces ratios of gamma functions by rising polynomials.

Rather than tire the skeptical reader with more preliminaries, it is perhaps best to move on to our examples without delay. The next section defines the MM principle, discusses our three driving inequalities, and reviews two simple acceleration methods. Section 3 derives MM algorithms for some standard multivariate discrete distributions, namely the Dirichlet-multinomial and Connor–Mosimann distributions, the Neerchal–Morel distribution, the negative-multinomial distribution, certain distributions on partitions, and zero-truncated and zero-inflated distributions. Section 4 describes a numerical experiment comparing the performance of the MM algorithms, accelerated MM algorithms, and Newton's method on model fitting of handwritten digit data. Our discussion concludes by mentioning directions for further research and by frankly acknowledging the limitations of the MM principle.

2. OVERVIEW OF THE MM ALGORITHM

As we have already emphasized, the MM algorithm is a principle for creating algorithms rather than a single algorithm. There are two versions of the MM principle, one for iterative minimization and another for iterative maximization. Here we deal only with the maximization version. Let $f(\theta)$ be the objective function we seek to maximize. An

MM algorithm involves minorizing $f(\theta)$ by a surrogate function $g(\theta|\theta^n)$ anchored at the current iterate θ^n of a search. Minorization is defined by the two properties

$$f(\theta^n) = g(\theta^n|\theta^n), \tag{2.1}$$

$$f(\theta) \geq g(\theta|\theta^n), \qquad \theta \neq \theta^n. \tag{2.2}$$

In other words, the surface $\theta \mapsto g(\theta|\theta^n)$ lies below the surface $\theta \mapsto f(\theta)$ and is tangent to it at the point $\theta = \theta^n$. Construction of the surrogate function $g(\theta|\theta^n)$ constitutes the first M of the MM algorithm.

In the second M of the algorithm, we maximize the surrogate $g(\theta|\theta^n)$ rather than $f(\theta)$. If θ^{n+1} denotes the maximum point of $g(\theta|\theta^n)$, then this action forces the ascent property $f(\theta^{n+1}) \geq f(\theta^n)$. The straightforward proof

$$f(\theta^{n+1}) \geq g(\theta^{n+1}|\theta^n) \geq g(\theta^n|\theta^n) = f(\theta^n)$$

reflects definitions (2.1) and (2.2) and the choice of θ^{n+1}. The ascent property is the source of the MM algorithm's numerical stability. Strictly speaking, it depends only on increasing $g(\theta|\theta^n)$, not on maximizing $g(\theta|\theta^n)$.

The art in devising an MM algorithm revolves around intelligent choices of minorizing functions. This brings us to the first of our three basic minorizations

$$\ln\left(\sum_{i=1}^m \alpha_i\right) \geq \sum_{i=1}^m \frac{\alpha_i^n}{\sum_{j=1}^m \alpha_j^n} \ln\left(\frac{\sum_{j=1}^m \alpha_j^n}{\alpha_i^n}\alpha_i\right), \tag{2.3}$$

invoking the chord below the graph property of the concave function $\ln x$. Note here that all parameter values are positive and that equality obtains whenever $\alpha_i = \alpha_i^n$ for all i. Our second basic minorization

$$-\ln(c+\alpha) \geq -\ln(c+\alpha^n) - \frac{1}{c+\alpha^n}(\alpha - \alpha^n) \tag{2.4}$$

restates the supporting hyperplane property of the convex function $-\ln(c+x)$. Our final basic minorization

$$-\ln(1-\alpha) \geq -\ln(1-\alpha^n) + \frac{\alpha^n}{1-\alpha^n}\ln\left(\frac{\alpha}{\alpha^n}\right) \tag{2.5}$$

is just a rearrangement of the two-point information inequality

$$\alpha^n \ln \alpha^n + (1-\alpha^n)\ln(1-\alpha^n) \geq \alpha^n \ln \alpha + (1-\alpha^n)\ln(1-\alpha).$$

Here α and α^n must lie in $(0, 1)$. Any standard text on inequalities, for example, the book by Steele (2004), proves these three inequalities. Because piecemeal minorization works well, our derivations apply the basic minorizations only to strategic parts of the overall objective function, leaving other parts untouched.

The convergence theory of MM algorithms is well known (Lange 2004). Convergence to a stationary point is guaranteed provided five properties of the objective function $f(\theta)$ and the MM algorithm map $M(\theta)$ hold: (a) $f(\theta)$ is coercive on its open domain; (b) $f(\theta)$ has only isolated stationary points; (c) $M(\theta)$ is continuous; (d) θ^* is a fixed point of $M(\theta)$ if and only if it is a stationary point of $f(\theta)$; (e) $f[M(\theta^*)] \leq f(\theta^*)$, with equality if and

only if θ^* is a fixed point of $M(\theta)$. Most of these conditions are easy to verify for our examples, so the details will be omitted.

A common criticism of EM and MM algorithms is their slow convergence. Fortunately, MM algorithms can be easily accelerated (Jamshidian and Jennrich 1995; Lange 1995a; Jamshidian and Jennrich 1997; Varadhan and Rolland 2008). We will employ two versions of the recent square iterative method (SQUAREM) developed by Varadhan and Roland (2008). These simple vector extrapolation techniques require computation of two MM updates at each iteration. Denote the two updates by $M(\theta^n)$ and $M \circ M(\theta^n)$, where $M(\theta)$ is the MM algorithm map. These updates in turn define two vectors

$$u = M(\theta^n) - \theta^n, \qquad v = M \circ M(\theta^n) - M(\theta^n) - u.$$

The versions diverge in how they compute the steplength constant s. SqMPE1 (minimal polynomial extrapolation) takes $s = \frac{u^t u}{u^t v}$, while SqRRE1 (reduced rank extrapolation) takes $s = \frac{u^t v}{v^t v}$. Once s is specified, we define the next accelerated iterate by $\theta^{n+1} = \theta^n - 2su + s^2 v$. Readers should consult the original article for motivation of SQUAREM. Whenever θ^{n+1} decreases the log-likelihood $L(\theta)$, we revert to the MM update $\theta^{n+1} = M \circ M(\theta^n)$. Finally, we declare convergence when

$$\frac{|L(\theta^n) - L(\theta^{n-1})|}{|L(\theta^{n-1})| + 1} < \epsilon. \tag{2.6}$$

In the numerical examples that follow, we use the stringent criterion $\epsilon = 10^{-9}$. More sophisticated stopping criteria based on the gradient of the objective function and the norm of the parameter increment lead to similar results.

3. APPLICATIONS

3.1 DIRICHLET-MULTINOMIAL AND CONNOR–MOSIMANN DISTRIBUTIONS

When count data exhibit overdispersion, the Dirichlet-multinomial distribution is often substituted for the multinomial distribution. The multinomial distribution is characterized by a vector $\mathbf{p} = (p_1, \ldots, p_d)$ of cell probabilities and a total number of trials m. In the Dirichlet-multinomial sampling, $\mathbf{p}$ is first drawn from a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_d)$. Once the cell probabilities are determined, multinomial sampling commences. This leads to the admixture density

$$\begin{aligned} h(\mathbf{x}|\alpha) &= \frac{\Gamma(|\alpha|)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_d)} \int_{\Delta_d} \binom{m}{\mathbf{x}} \prod_{i=1}^{d} p_i^{x_i + \alpha_i - 1} \, dp_1 \cdots dp_d \\ &= \binom{m}{\mathbf{x}} \frac{\Gamma(\alpha_1 + x_1) \cdots \Gamma(\alpha_d + x_d)}{\Gamma(|\alpha| + m)} \frac{\Gamma(|\alpha|)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_d)}, \end{aligned} \tag{3.1}$$

where $|\alpha| = \sum_{i=1}^{d} \alpha_i$, Δ_d is the unit simplex in $\mathbb{R}^d$, and $\mathbf{x} = (x_1, \ldots, x_d)$ is the vector of cell counts. Note that the count total $|x| = \sum_{i=1}^{d} x_i$ is fixed at m. Standard calculations

show that a random vector $\mathbf{X}$ drawn from $h(\mathbf{x}|\alpha)$ has the means, variances, and covariances

$$\begin{aligned}
\mathbf{E}(\mathbf{X}_i) &= m\frac{\alpha_i}{|\alpha|},\\
\mathbf{Var}(\mathbf{X}_i) &= m\frac{\alpha_i}{|\alpha|}\left(1-\frac{\alpha_i}{|\alpha|}\right)\frac{|\alpha|+m}{|\alpha|+1},\\
\mathbf{Cov}(\mathbf{X}_i,\mathbf{X}_j) &= -m\frac{\alpha_i}{|\alpha|}\frac{\alpha_j}{|\alpha|}\frac{|\alpha|+m}{|\alpha|+1}, \qquad i\neq j.
\end{aligned}$$

If the fractions $\frac{\alpha_i}{|\alpha|}$ tend to constants p_i as $|\alpha|$ tends to ∞, then these moments collapse to the corresponding moments of the multinomial distribution with proportions $p_1,\ldots,p_d$.

One of the most unappealing features of the density function $h(\mathbf{x}|\alpha)$ is the occurrence of the gamma function. Fortunately, very early on Haldane (1941) noted the alternative representation

$$h(\mathbf{x}|\alpha)=\binom{m}{\mathbf{x}}\frac{\prod_{j=1}^d\alpha_j(\alpha_j+1)\cdots(\alpha_j+x_j-1)}{|\alpha|(|\alpha|+1)\cdots(|\alpha|+m-1)}. \tag{3.2}$$

The replacement of gamma functions by rising polynomials is a considerable gain in simplicity. Bailey (1957) later suggested the reparameterization

$$\pi_j=\frac{\alpha_j}{|\alpha|}, \qquad j=1,\ldots,d, \qquad \theta=\frac{1}{|\alpha|}$$

in terms of the proportion vector $\pi=(\pi_1,\ldots,\pi_d)$ and the overdispersion parameter θ. In this setting, the discrete density function becomes

$$h(\mathbf{x}|\pi,\theta)=\binom{m}{\mathbf{x}}\frac{\prod_{j=1}^d\pi_j(\pi_j+\theta)\cdots[\pi_j+(x_j-1)\theta]}{(1+\theta)\cdots[1+(m-1)\theta]}. \tag{3.3}$$

This version of the density function is used to good effect by Griffiths (1973) in implementing Newton's method for maximum likelihood estimation with the beta-binomial distribution.

In maximum likelihood estimation, we pass to log-likelihoods. This introduces logarithms and turns factors into sums. To construct an MM algorithm under the parameterization (3.2), we need to minorize terms such as $\ln(\alpha_j+k)$ and $-\ln(|\alpha|+k)$. The basic inequalities (2.3) and (2.4) are directly relevant. Suppose we draw t independent samples $\mathbf{x}_1,\ldots,\mathbf{x}_t$ from the Dirichlet-multinomial distribution with m_i trials for sample i. The term $-\ln(|\alpha|+k)$ occurs in the log-likelihood for $\mathbf{x}_i$ if and only if $m_i\geq k+1$. Likewise the term $\ln(\alpha_j+k)$ occurs in the log-likelihood for $\mathbf{x}_i$ if and only if $x_{ij}\geq k+1$. It follows that the log-likelihood for the entire sample can be written as

$$\begin{aligned}
L(\alpha) &= -\sum_k r_k\ln(|\alpha|+k)+\sum_j\sum_k s_{jk}\ln(\alpha_j+k),\\
r_k &= \sum_i 1_{\{m_i\geq k+1\}}, \qquad s_{jk}=\sum_i 1_{\{x_{ij}\geq k+1\}}.
\end{aligned} \tag{3.4}$$

The index k in these formulas ranges from 0 to $\max_i m_i-1$.

Applying our two basic minorizations to $L(\alpha)$ yields the surrogate function

$$g(\alpha|\alpha^n) = -\sum_k r_k \frac{1}{|\alpha^n| + k}|\alpha| + \sum_j \sum_k s_{jk} \frac{\alpha_j^n}{\alpha_j^n + k} \ln \alpha_j$$

up to an irrelevant additive constant. Equating the partial derivative of the surrogate with respect to α_j to 0 produces the simple MM update

$$\alpha_j^{n+1} = \left(\sum_k \frac{s_{jk}\alpha_j^n}{\alpha_j^n + k}\right) \Big/ \left(\sum_k \frac{r_k}{|\alpha^n| + k}\right). \tag{3.5}$$

Minka (2003) derived these updates from a different perspective.

Under the parameterization (3.3), matters are slightly more complicated. Now we minorize the terms $-\ln(1 + k\theta)$ and $\ln(\pi_j + k\theta)$ via

$$-\log(1 + k\theta) \geq -\log(1 + k\theta^n) - \frac{1}{1 + k\theta^n}(k\theta - k\theta^n)$$

and

$$\log(\pi_j + k\theta) \geq \frac{\pi_j^n}{\pi_j^n + k\theta^n} \log\left(\frac{\pi_j^n + k\theta^n}{\pi_j^n}\pi_j\right) + \frac{k\theta^n}{\pi_j^n + k\theta^n} \log\left(\frac{\pi_j^n + k\theta^n}{k\theta^n}k\theta\right).$$

These minorizations lead to the surrogate function

$$-\sum_k r_k \frac{k}{1 + k\theta^n}\theta + \sum_j \sum_k s_{jk}\left\{\frac{\pi_j^n}{\pi_j^n + k\theta^n} \log \pi_j + \frac{k\theta^n}{\pi_j^n + k\theta^n} \log \theta\right\}$$

up to an irrelevant constant. Setting the partial derivative with respect to θ equal to 0 yields the MM update

$$\theta^{n+1} = \left(\sum_j \sum_k \frac{s_{jk}k\theta^n}{\pi_j^n + k\theta^n}\right) \Big/ \left(\sum_k \frac{r_k k}{1 + k\theta^n}\right). \tag{3.6}$$

The update of the proportion vector π must be treated as a Lagrange multiplier problem owing to the constraint $\sum_j \pi_j = 1$. Familiar arguments produce the MM update

$$\pi_j^{n+1} = \left(\sum_k \frac{s_{jk}\pi_j^n}{\pi_j^n + k\theta^n}\right) \Big/ \left(\sum_l \sum_k \frac{s_{lk}\pi_l^n}{\pi_l^n + k\theta^n}\right). \tag{3.7}$$

The two updates summarized by (3.5), (3.6), and (3.7) enjoy several desirable properties. First, parameter constraints are built in. Second, stationary points of the log-likelihood are fixed points of the updates. Virtually all MM algorithms share these properties. The update (3.7) also reduces to the maximum likelihood estimate

$$\hat{\pi}_j = \frac{\sum_k s_{jk}}{\sum_l \sum_k s_{lk}} = \frac{\sum_i x_{ij}}{\sum_i m_i} \tag{3.8}$$

of the corresponding multinomial proportion when $\theta^n = 0$.

The estimate (3.8) furnishes a natural initial value π_j^0. To derive an initial value for the overdispersion parameter θ, consider the first two moments

$$\mathbf{E}(P_j) = \frac{\alpha_j}{|\alpha|}, \qquad \mathbf{E}(P_j^2) = \frac{\alpha_j(\alpha_j + 1)}{|\alpha|(|\alpha| + 1)}$$

of a Dirichlet distribution with parameter vector α. These identities imply that

$$\sum_{j=1}^{d} \frac{\mathbf{E}(P_j^2)}{\mathbf{E}(P_j)} = \frac{|\alpha| + d}{|\alpha| + 1} = \rho,$$

which can be solved for $\theta = 1/|\alpha|$ in terms of ρ as $\theta = (\rho - 1)/(d - \rho)$. Substituting the estimate

$$\hat{\rho} = \sum_j \frac{\sum_i (x_{ij}/m_i)^2}{\sum_i (x_{ij}/m_i)}$$

for ρ gives a sensible initial value θ^0.

To test our two MM algorithms, we now turn to the beta-binomial data of Haseman and Soares (1976) on male mice exposed to various mutagens. The two outcome categories are (a) dead implants and (b) survived implants. In their first dataset, there are $t = 524$ observations with between $m = 1$ and $m = 20$ trials per observation. Table 1 presents the final log-likelihood, number of iterations, and running time (in seconds) of the two MM algorithms and their SQUAREM accelerations on these data. All MM algorithms converge to the maximum point previously found by the scoring method (Paul, Balasooriya, and Banerjee 2005). For the choice $\epsilon = 10^{-9}$ in stopping criterion (2.6), the MM algorithm (3.5) takes 700 iterations and 0.1580 sec to converge on a laptop computer. The alternative MM algorithm given in the updates (3.6) and (3.7) takes 339 iterations and 0.1626 sec. Figure 1 depicts the progress of the MM iterates on a contour plot of the log-likelihood. The conventional MM algorithm crawls slowly along the ridge in the contour plot; the accelerated versions SqMPE1 and SqRRE1 significantly reduce both the number of iterations and the running time until convergence.

The Dirichlet-multinomial distribution suffers from two restrictions that limit its applicability, namely the negative correlation of coordinates and the determination of variances by means. It is possible to overcome these restrictions by choosing a more flexible mixing distribution as a prior for the multinomial. Connor and Mosimann (1969) suggested a generalization of the Dirichlet distribution that meets this challenge. The resulting

Table 1. MM algorithms for the Haseman and Soares beta-binomial data.

	(π, θ) parameterization			α parameterization		
Algorithm	L	# iters	Time	L	# iters	Time
MM	−777.79	339	0.1626	−777.79	700	0.1580
SqMPE1 MM	−777.79	10	0.0093	−777.79	14	0.0105
SqRRE1 MM	−777.79	18	0.0159	−777.79	14	0.0100

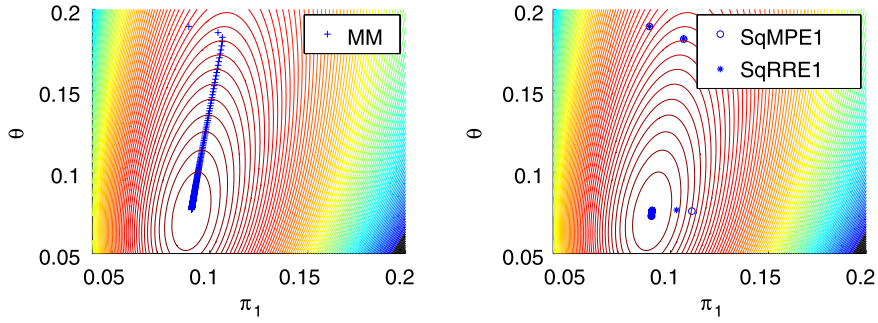

Figure 1. MM Ascent of the Dirichlet-multinomial log-likelihood surface. A color version of this figure is available in the electronic version of this article.

admixed distribution, called the generalized Dirichlet-multinomial distribution, has proved its worth in machine learning problems such as the modeling and clustering of images, handwritten digits, and text documents (Bouguila 2008). It is therefore helpful to derive an MM algorithm for maximum likelihood estimation with this distribution that avoids the complications of gamma/digamma/trigamma functions arising with Newton's method (Bouguila 2008). The Connor–Mosimann distribution is constructed inductively by the mechanism of stick breaking. Imagine breaking the interval $[0, 1]$ into d subintervals of lengths $P_1, \ldots, P_d$ by choosing $d-1$ independent beta variates Z_i with parameters α_i and β_i. The length of subinterval 1 is $P_1 = Z_1$. Given P_1 through P_i, the length of subinterval $i+1$ is $P_{i+1} = Z_{i+1}(1 - P_1 - \cdots - P_i)$. The last length $P_d = 1 - (P_1 + \cdots + P_{d-1})$ takes up the slack. Standard calculations show that the P_i have the joint density

$$g(\mathbf{p}|\alpha, \beta) = \prod_{j=1}^{d-1} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} p_j^{\alpha_j - 1} \left(1 - \sum_{i=1}^{j} p_i\right)^{\gamma_j}, \qquad \mathbf{p} \in \Delta_d,$$

where $\gamma_j = \beta_j - \alpha_{j+1} - \beta_{j+1}$ for $j = 1, \ldots, d-2$ and $\gamma_{d-1} = \beta_{d-1} - 1$. The univariate case ($d = 2$) corresponds to the beta distribution. The Dirichlet distribution is recovered by taking $\beta_j = \alpha_{j+1} + \cdots + \alpha_d$. With $d-2$ more parameters than the Dirichlet distribution, the Connor–Mosimann distribution is naturally more versatile.

The Connor–Mosimann distribution is again conjugate to the multinomial distribution, and the marginal density of a count vector $\mathbf{X}$ over m trials is easily shown to be

$$\begin{aligned} \Pr(\mathbf{X} = \mathbf{x}) &= \int_{\Delta_d} \binom{m}{\mathbf{x}} \prod_{j=1}^{d-1} p_j^{x_j} g(\mathbf{p}|\alpha, \beta)\, d\mathbf{p} \\ &= \binom{m}{\mathbf{x}} \prod_{j=1}^{d-1} \frac{\Gamma(\alpha_j + x_j)}{\Gamma(\alpha_j)} \frac{\Gamma(\beta_j + y_{j+1})}{\Gamma(\beta_j)} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j + \beta_j + y_j)}, \end{aligned}$$

where $y_j = \sum_{k=j}^{d} x_k$. If we adopt the reparameterization

$$\theta_j = \frac{1}{\alpha_j + \beta_j}, \qquad \pi_j = \frac{\alpha_j}{\alpha_j + \beta_j}, \qquad j = 1, \ldots, d-1,$$

and use the fact that $x_j + y_{j+1} = y_j$, then the density can be re-expressed as

$$\binom{m}{\mathbf{x}} \prod_{j=1}^{d-1} \frac{\pi_j \cdots [\pi_j + (x_j - 1)\theta_j] \times (1-\pi_j) \cdots [1 - \pi_j + (y_{j+1} - 1)\theta_j]}{1 \cdots [1 + (y_j - 1)\theta_j]}. \quad (3.9)$$

Thus, maximum likelihood estimation of the parameter vectors $\pi = (\pi_1, \ldots, \pi_{d-1})$ and $\theta = (\theta_1, \ldots, \theta_{d-1})$ by the MM algorithm reduces to the case of $d-1$ independent beta-binomial problems.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_t$ be a random sample from the generalized Dirichlet-multinomial distribution (3.9) with m_i trials for observation $\mathbf{x}_i$. Following our reasoning for estimation with the Dirichlet-multinomial, we define the associated counts

$$r_{jk} = \sum_{i=1}^{t} 1_{\{x_{ij} \geq k+1\}}, \qquad s_{jk} = \sum_{i=1}^{t} 1_{\{y_{ij} \geq k+1\}}$$

for $1 \leq j \leq d-1$. In this notation, the reader can readily check that the MM updates become

$$\pi_j^{n+1} = \left(\sum_k \frac{r_{jk}\pi_j^n}{\pi_j^n + k\theta_j^n}\right) \Big/ \left(\sum_k \left[\frac{r_{jk}\pi_j^n}{\pi_j^n + k\theta_j^n} + \frac{s_{j+1,k}(1-\pi_j^n)}{1 - \pi_j^n + k\theta_j^n}\right]\right),$$

$$\theta_j^{n+1} = \left(\sum_k \left[\frac{r_{jk}k\theta_j^n}{\pi_j^n + k\theta_j^n} + \frac{s_{j+1,k}k\theta_j^n}{1 - \pi_j^n + k\theta_j^n}\right]\right) \Big/ \left(\sum_k \frac{s_{jk}}{1 + k\theta_j^n}\right).$$

3.2 NEERCHAL–MOREL DISTRIBUTION

Neerchal and Morel (1998, 2005) proposed an alternative to the Dirichlet-multinomial distribution that accounts for overdispersion by finite admixture. If $\mathbf{x}$ represents count data over m trials and d categories, then their discrete density is

$$h(\mathbf{x}|\pi, \rho) = \sum_{j=1}^{d} \pi_j \binom{m}{\mathbf{x}} [(1-\rho)\pi_1]^{x_1} \cdots [(1-\rho)\pi_j + \rho]^{x_j} \cdots [(1-\rho)\pi_d]^{x_d}, \quad (3.10)$$

where $\pi = (\pi_1, \ldots, \pi_d)$ is a vector of proportions and $\rho \in [0, 1]$ is an overdispersion parameter. The Neerchal–Morel distribution collapses to the multinomial distribution when $\rho = 0$. Straightforward calculations show that the Neerchal–Morel distribution has means, variances, and covariances

$$\begin{aligned} \mathbf{E}(\mathbf{X}_i) &= m\pi_i, \\ \mathbf{Var}(\mathbf{X}_i) &= m\pi_i(1-\pi_i)[1 - \rho^2 + m\rho^2], \\ \mathbf{Cov}(\mathbf{X}_i, \mathbf{X}_j) &= -m\pi_i\pi_j[1 - \rho^2 + m\rho^2], \qquad i \neq j. \end{aligned}$$

These are precisely the same as the first- and second-order moments of the Dirichlet-multinomial distribution provided we identify $\pi_i = \alpha_i/|\alpha|$ and $\rho^2 = 1/(|\alpha| + 1)$.

If we draw t independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_t$ from the Neerchal–Morel distribution with m_i trials for sample i, then the log-likelihood is

$$\sum_i \ln\left\{\sum_j \pi_j \binom{m_i}{\mathbf{x}_i} [(1-\rho)\pi_1]^{x_{i1}} \cdots [(1-\rho)\pi_j + \rho]^{x_{ij}} \cdots [(1-\rho)\pi_d]^{x_{id}}\right\}. \quad (3.11)$$

It is worth bearing in mind that every mixture model yields to the minorization (2.3). This is one of the secrets to the success of the EM algorithm. As a practical matter, explicit minorization via inequality (2.3) is more mechanical and often simpler to implement than performing the E step of the EM algorithm. This is particularly true when several minorizations intervene before we reach the ideal surrogate. Here two successive minorizations are needed.

To state the first minorization, let us abbreviate

$$\Pi_{ij} = \pi_j [(1-\rho)\pi_1]^{x_{i1}} \cdots [(1-\rho)\pi_j + \rho]^{x_{ij}} \cdots [(1-\rho)\pi_d]^{x_{id}}$$

and denote by Π_{ij}^n the same quantity evaluated at the nth iterate. In this notation it follows that

$$\ln\left(\sum_j \Pi_{ij}\right) \geq \sum_j w_{ij}^n \ln\left(\frac{\Pi_{ij}}{w_{ij}^n}\right) = \sum_j w_{ij}^n \ln \Pi_{ij} - \sum_j w_{ij}^n \ln w_{ij}^n$$

with weights $w_{ij}^n = \frac{\Pi_{ij}^n}{\sum_l \Pi_{il}^n}$. The logarithm splits $\ln \Pi_{ij}$ into the sum

$$\ln \Pi_{ij} = m_i \ln(1-\rho) + \ln \pi_j + x_{i1} \ln \pi_1 + \cdots + x_{ij} \ln(\pi_j + \theta) + \cdots + x_{id} \ln \pi_d$$

for $\theta = \rho/(1-\rho)$. To separate the parameters π_j and θ in the troublesome term $\ln(\pi_j + \theta)$, we apply the minorization (2.3) again. This produces

$$\ln(\pi_j + \theta) \geq \frac{\pi_j^n}{\pi_j^n + \theta^n} \ln\left(\frac{\pi_j^n + \theta^n}{\pi_j^n}\pi_j\right) + \frac{\theta^n}{\pi_j^n + \theta^n} \ln\left(\frac{\pi_j^n + \theta^n}{\theta^n}\theta\right),$$

and up to a constant the surrogate function takes the form

$$\sum_i \sum_j w_{ij}^n \left[\sum_k x_{ik} \ln \pi_k + \left(1 - \frac{x_{ij}\theta^n}{\pi_j^n + \theta^n}\right) \ln \pi_j\right]$$
$$+ \sum_i \sum_j w_{ij}^n \left[\left(m_i - \frac{x_{ij}\theta^n}{\pi_j^n + \theta^n}\right) \ln(1-\rho) + \frac{x_{ij}\theta^n}{\pi_j^n + \theta^n} \ln \rho\right].$$

Standard arguments now yield the updates

$$\pi_k^{n+1} = \left(\sum_i \sum_j w_{ij}^n x_{ik} + \sum_i w_{ik}^n \left(1 - \frac{x_{ik}\theta^n}{\pi_k^n + \theta^n}\right)\right)$$
$$\Big/ \left(\sum_l \sum_i \sum_j w_{ij}^n x_{il} + \sum_l \sum_i w_{il}^n \left(1 - \frac{x_{il}\theta^n}{\pi_l^n + \theta^n}\right)\right),$$

$$\rho^{n+1} = \left(\sum_i \sum_j \frac{w_{ij}^n x_{ij}\theta^n}{\pi_j^n + \theta^n}\right) \Big/ \left(\sum_i m_i\right), \qquad \theta^{n+1} = \frac{\rho^{n+1}}{1 - \rho^{n+1}}.$$

Table 2. Performance of the Neerchal–Morel MM algorithms.

Algorithm	L	# iters	Time
MM	−783.29	128	0.2289
SqMPE1 MM	−783.29	10	0.0207
SqRRE1 MM	−783.29	11	0.0221

Table 2 lists convergence results for this MM algorithm and its SQUAREM accelerations on the previously discussed Haseman and Soares data.

3.3 NEGATIVE-MULTINOMIAL

The motivation for the negative-multinomial distribution comes from multinomial sampling with $d+1$ categories assigned probabilities $\pi_1, \ldots, \pi_{d+1}$. Sampling continues until category $d+1$ accumulates β outcomes. At that moment we count the number of outcomes x_i falling in category i for $1 \le i \le d$. For a given vector $\mathbf{x} = (x_1, \ldots, x_d)$, elementary combinatorics gives the probability

$$\begin{aligned} h(\mathbf{x}|\beta, \pi) &= \binom{\beta + |\mathbf{x}| - 1}{|\mathbf{x}|} \binom{|\mathbf{x}|}{\mathbf{x}} \prod_{i=1}^{d} \pi_i^{x_i} \pi_{d+1}^{\beta} \\ &= \frac{\beta(\beta+1)\cdots(\beta+|\mathbf{x}|-1)}{x_1! \cdots x_d!} \prod_{i=1}^{d} \pi_i^{x_i} \pi_{d+1}^{\beta}. \end{aligned} \tag{3.12}$$

This formula continues to make sense even if the positive parameter β is not an integer. For arbitrary $\beta > 0$, the most straightforward way to construct the negative-multinomial distribution is to run d independent Poisson processes with intensities $\pi_1, \ldots, \pi_d$. Wait a gamma distributed length of time with shape parameter β and intensity parameter π_{d+1}. At the expiration of this waiting time, count the number of random events X_i of each type i among the first d categories. The random vector $\mathbf{X}$ has precisely the discrete density (3.12).

The Poisson process perspective readily yields the moments

$$\begin{aligned} \mathbf{E}(X_i) &= \beta \frac{\pi_i}{\pi_{d+1}}, \\ \mathbf{Var}(X_i) &= \beta \frac{\pi_i}{\pi_{d+1}} \left(1 + \frac{\pi_i}{\pi_{d+1}}\right), \\ \mathbf{Cov}(X_i, X_j) &= \beta \frac{\pi_i}{\pi_{d+1}} \frac{\pi_j}{\pi_{d+1}}, \qquad i \neq j. \end{aligned} \tag{3.13}$$

Compared to a Poisson distributed random variable with the same mean, the component X_i is overdispersed. Also in contrast to the multinomial and Dirichlet-multinomial distributions, the counts from a negative-multinomial are positively correlated. Negative-multinomial sampling is therefore appealing in many applications.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_t$ be a random sample from the negative-multinomial distribution with $m_i = |\mathbf{x}_i|$. To maximize the log-likelihood

$$L(\beta, \pi) = \sum_k r_k \ln(\beta + k) + \sum_{j=1}^d x_{\cdot j} \ln \pi_j + t\beta \ln \pi_{d+1} - \sum_i \sum_j \ln x_{ij}!,$$
$$r_k = \sum_i 1_{\{m_i \geq k+1\}}, \qquad x_{\cdot j} = \sum_i x_{ij},$$

we must deal with the terms $\ln(\beta + k)$. Fortunately, the minorization (2.4) implies

$$\ln(\beta + k) \geq \frac{\beta^n}{\beta^n + k} \ln\left(\frac{\beta^n + k}{\beta^n}\beta\right) + \frac{k}{\beta^n + k} \ln\left(\frac{\beta^n + k}{k}k\right),$$

leading to the surrogate function

$$g(\beta, \pi | \beta^n, \pi^n) = \sum_k r_k \frac{\beta^n}{\beta^n + k} \ln \beta + \sum_{j=1}^d x_{\cdot j} \ln \pi_j + t\beta \ln \pi_{d+1}$$

up to an irrelevant constant. In view of the constraint $\pi_{d+1} = 1 - \sum_{j=1}^d \pi_j$, the stationarity conditions for a maximum of the surrogate reduce to

$$0 = \frac{1}{\beta} \sum_k r_k \frac{\beta^n}{\beta^n + k} + t \ln \pi_{d+1}, \qquad 0 = \frac{x_{\cdot j}}{\pi_j} - \frac{t\beta}{\pi_{d+1}}, \qquad 1 \leq j \leq d. \tag{3.14}$$

Unfortunately, it is impossible to solve this system of equations analytically. There are two resolutions to the dilemma. One is block relaxation (de Leeuw 1994) alternating the updates

$$\beta^{n+1} = -\left(\sum_k r_k \frac{\beta^n}{\beta^n + k}\right) \Big/ (t \ln \pi_{d+1}^n)$$

and

$$\pi_{d+1}^{n+1} = \frac{t\beta^{n+1}}{\sum_{k=1}^d x_{\cdot k} + t\beta^{n+1}}, \qquad \pi_j^{n+1} = \frac{x_{\cdot j}}{\sum_{k=1}^d x_{\cdot k} + t\beta^{n+1}}, \qquad 1 \leq j \leq d.$$

This strategy enjoys the ascent property of all MM algorithms.

The other possibility is to solve the stationarity equations numerically. It is clear that the system of equations (3.14) reduces to the single equation

$$0 = \frac{1}{\beta} \sum_k r_k \frac{\beta^n}{\beta^n + k} + t \ln\left(\frac{t\beta}{\sum_{k=1}^d x_{\cdot k} + t\beta}\right)$$

for β. Equivalently, if we let

$$\alpha = \beta^{-1}, \qquad \bar{m} = \frac{1}{t} \sum_i \sum_j x_{ij}, \qquad c^n = \sum_k r_k \frac{\beta^n}{\beta^n + k},$$

then we must find a root of the equation $f(\alpha) = \alpha c^n - t \ln(\alpha \bar{m} + 1) = 0$. It is clear that $f(\alpha)$ is a strictly convex function with $f(0) = 0$ and $\lim_{\alpha \to \infty} f(\alpha) = \infty$. Furthermore, a little reflection shows that $f'(0) = c^n - t\bar{m} < 0$. Thus, there is a single root of $f(\alpha)$ on the

interval $(0, \infty)$. Owing to the convexity of $f(\alpha)$, Newton's method will reliably find the root if started to the right of the minimum of $f(\alpha)$ at $\alpha = t/c^n - 1/\bar{m}$.

To find initial values, we again resort to the method of moments. Based on the moments (3.13), the mean and variance of $|\mathbf{X}| = \sum_k X_j$ are

$$\mathbf{E}(|\mathbf{X}|) = \frac{\beta(1-\pi_{d+1})}{\pi_{d+1}}, \qquad \mathbf{Var}(|\mathbf{X}|) = \frac{\beta(1-\pi_{d+1})}{\pi_{d+1}^2}.$$

These suggest that we take

$$\beta^0 = \frac{\bar{x}^2}{s^2 - \bar{x}}, \qquad \pi_{d+1}^0 = \frac{\bar{x}}{s^2}, \qquad \pi_j^0 = \frac{\pi_{d+1}^0}{\beta^0}\frac{x_{\cdot j}}{t}, \qquad 1 \le j \le d,$$

where

$$\bar{x} = \frac{1}{t}\sum_{i=1}^{t} m_i, \qquad s^2 = \frac{1}{t-1}\sum_{i=1}^{t}(m_i - \bar{x})^2.$$

When the data are underdispersed ($s^2 < \bar{x}$), our proposed initial values are not meaningful, but a negative-multinomial model is a poor choice anyway.

3.4 DISTRIBUTIONS ON PARTITIONS

A partition of a positive integer m into k parts is a vector $\mathbf{a} = (a_1, \ldots, a_m)$ of nonnegative integers such that $\sum_i a_i = k$ and $|\mathbf{a}| = \sum_i i a_i = m$. In population genetics, the partition distributions of Ewens (2004) and Pitman (Pitman 1995; Johnson, Kotz, and Balakrishnan 1997) find wide application. We now develop an MM algorithm for Pitman's distribution, which generalizes Ewens's distribution. Pitman's distribution

$$\begin{aligned}\Pr(\mathbf{A} = \mathbf{a}|m, \alpha, \theta) = {} & \frac{m! \prod_{i=1}^{|\mathbf{a}|-1}(\theta + i\alpha)}{(\theta+1)\cdots(\theta+m-1)} \\ & \times \prod_{j=1}^{m}\left[\frac{(1-\alpha)\cdots(1-\alpha+j-2)]}{i!}\right]^{a_j}\frac{1}{a_j!}\end{aligned}$$

involves two parameters $0 \le \alpha < 1$ and $\theta > -\alpha$. Ewens's distribution corresponds to the choice $\alpha = 0$. We will restrict θ to be positive.

To estimate parameters given u independent partitions $\mathbf{a}_1, \ldots, \mathbf{a}_u$ from Pitman's distribution, we use the minorizations (2.3) and (2.4) to derive the minorizations

$$\begin{aligned}\ln(\theta + i\alpha) &\ge \frac{\theta^n}{\theta^n + i\alpha^n}\ln\theta + \frac{i\alpha^n}{\theta^n + i\alpha^n}\ln\alpha + c, \\ \ln(1 - \alpha + i) &\ge \frac{1-\alpha^n}{1-\alpha^n+i}\ln(1-\alpha) + c, \\ -\ln(\theta + i) &\ge -\frac{1}{\theta^n + i}\theta + c,\end{aligned}$$

where c is a different irrelevant constant in each case. Assuming $\mathbf{a}_j$ is a partition of the integer m_j, it follows that the log-likelihood is minorized by

$$\sum_i \frac{r_i\theta^n}{\theta^n + i\alpha^n} \ln\theta + \sum_i \frac{r_i i\alpha^n}{\theta^n + i\alpha^n} \ln\alpha + \sum_i \frac{s_i(1-\alpha^n)}{1-\alpha^n+i} \ln(1-\alpha) - \sum_i \frac{t_i}{\theta^n + i}\theta + c,$$

where

$$r_i = \sum_j 1_{\{|\mathbf{a}_j| \geq i+1\}}, \qquad s_i = \sum_j \sum_{k \geq i+2} a_{jk}, \qquad t_i = \sum_j 1_{\{m_i \geq i+1\}}.$$

Standard arguments now yield the simple updates

$$\alpha^{n+1} = \left(\sum_i \frac{r_i i\alpha^n}{\theta^n + i\alpha^n}\right) \Big/ \left(\sum_i \frac{r_i i\alpha^n}{\theta^n + i\alpha^n} + \sum_i \frac{s_i(1-\alpha^n)}{1-\alpha^n+i}\right),$$

$$\theta^{n+1} = \left(\sum_i \frac{r_i\theta^n}{\theta^n + i\alpha^n}\right) \Big/ \left(\sum_i \frac{t_i}{\theta^n + i}\right).$$

If we set $\alpha^0 = 0$, then in all subsequent iterates $\alpha^n = 0$, and we get the MM updates for Ewens's distribution. Despite the availability of the moments of the parts A_i (Charalambides 2007), it is not clear how to initialize α and θ. Unfortunately, the alternative suggestion of Nobuaki (2001) does not guarantee that the initial values satisfy the constraints $\alpha \in [0, 1)$ and $\theta > 0$.

3.5 ZERO-TRUNCATED AND ZERO-INFLATED DATA

In this section we briefly indicate how the MM perspective sheds fresh light on EM algorithms for zero-truncated and zero-inflated data. Once again mastery of a handful of inequalities rather than computation of conditional expectations drives the derivations.

In many discrete probability models, only data with positive counts are observed. Counts that are 0 are missing. If $f(x|\theta)$ represents the density of the complete data, then the density of a random sample $x_1, \ldots, x_t$ of zero-truncated data amounts to

$$h(x|\theta) = \prod_{i=1}^t \frac{f(x_i|\theta)}{1 - f(0|\theta)}.$$

Inequality (2.5) immediately implies the minorization

$$\ln h(x|\theta) \geq \sum_{i=1}^t \left[\ln f(x_i|\theta) + \frac{f(0|\theta^n)}{1 - f(0|\theta^n)} \ln f(0|\theta)\right] + c,$$

where c is an irrelevant constant. In many models, maximization of this surrogate function is straightforward.

For instance, with zero-truncated data from the binomial, Poisson, and negative-binomial distributions, the MM updates reduce to

$$p^{n+1} = \Big(\sum_i x_i\Big)\Big/\Big(\sum_i \frac{m_i}{1-(1-p^n)^{m_i}}\Big), \qquad \lambda^{n+1} = \Big(\sum_i x_i\Big)\Big/\Big(\sum_i \frac{1}{1-e^{-\lambda^n}}\Big),$$
$$p^{n+1} = \Big(\sum_i \frac{m_i}{1-(p^n)^{m_i}}\Big)\Big/\Big(\sum_i\Big[x_i + \frac{m_i}{1-(p^n)^{m_i}}\Big]\Big).$$

For observation i of the binomial model, there are x_i successes out of m_i trials with success probability p per trial. λ is the mean in the Poisson model. For observation i of the negative-binomial model, there are x_i failures before m_i required successes.

More complicated models can be handled in similar fashion. The key insight in each case is to augment every ordinary observation $x_i > 0$ by a total of $f(0|\theta^n)/[1-f(0|\theta^n)]$ pseudo-observations of 0 at iteration n. With this amendment, the two MM algorithms for the beta-binomial distribution implemented in (3.5), (3.6), and (3.7) remain valid except that the count variables r_k and s_{jk} defining the updated parameters at iteration n become

$$s_{1k} = \sum_i 1_{\{x_{i1}\geq k+1\}}, \qquad s_{2k} = \sum_i\Big[1_{\{x_{i2}\geq k+1\}} + \frac{f(0|\pi^n,\theta^n)}{1-f(0|\pi^n,\theta^n)}\Big],$$
$$r_k = \sum_i\Big[1 + \frac{f(0|\pi^n,\theta^n)}{1-f(0|\pi^n,\theta^n)}\Big]1_{\{m_i\geq k+1\}},$$

where

$$f(0|\pi^n,\theta^n) = \frac{\pi_2^n(\pi_2^n+\theta^n)\cdots[\pi_2^n+(m_i-1)\theta^n]}{(1+\theta^n)\cdots[1+(m_i-1)\theta^n]}.$$

Here category 1 represents success and category 2 failure. If we start with $\theta^0 = 0$, then we recover the updates for the zero-truncated binomial distribution.

Zero-inflated data are equally easy to handle. The density function is now

$$h(x|\theta,\pi) = \prod_{i=1}^t[(1-\pi)+\pi f(0|\theta)]^{1_{\{x_i=0\}}}[\pi f(x_i|\theta)]^{1_{\{x_i>0\}}}.$$

Inequality (2.3) entails the minorization

$$\ln h(x|\theta,\pi) \geq \sum_{i=1}^t 1_{\{x_i=0\}}\big\{z^n\ln(1-\pi)+(1-z_n)[\ln\pi+\ln f(0|\theta^n)]\big\}$$
$$+\sum_{i=1}^t 1_{\{x_i>0\}}[\ln\pi+\ln f(x_i|\theta)],$$
$$z^n = \frac{1-\pi^n}{1-\pi^n+\pi^n f(0|\theta^n)}.$$

The MM update of the inflation-admixture parameter clearly is

$$\pi^{n+1} = \frac{1}{t}\sum_{i=1}^t\big[1_{\{x_i>0\}} + 1_{\{x_i=0\}}(1-z^n)\big].$$

As a typical example, consider estimation with the zero-inflated Poisson (Patil 2007). The mean λ of the Poisson component is updated by

$$\lambda^{n+1} = \frac{\sum_i x_i}{\sum_i [(1 - z^n) 1_{\{x_i = 0\}} + 1_{\{x_i > 0\}}]}.$$

In other words, every 0 observation is discounted by the amount z^n at iteration n. This makes intuitive sense.

4. A NUMERICAL EXPERIMENT

As a numerical experiment, we fit the Dirichlet-multinomial (two parameterizations) and the Neerchal–Morel distributions to the 3823 training digits in the handwritten digit data from the UCI machine learning repository (Asuncion and Newman 2007). Each normalized 32×32 bitmap is divided into 64 blocks of size 4×4, and the black pixels are counted in each block. This generates a 64-dimensional count vector for each bitmap. Bouguila (2008) successfully fit mixtures of Connor–Mosimann to the training data and used the estimated models to cluster the test data. For illustrative purposes we now fit the Dirichlet-multinomial (two parameterizations) and Neerchal–Morel models. Based on the majorization (2.3), it is straightforward to extend our MM algorithms to fit finite mixture models using any of the previously encountered multivariate discrete distributions.

Table 3 lists the final log-likelihoods, number of iterations, and running times of the different algorithms tested. The MM and accelerated MM algorithms were coded in plain Matlab script language. Newton's method was implemented using the `fmincon` function in the Matlab Optimization Toolbox under the interior-point option with user-supplied analytical gradient and Hessian. All iterations started from the initial points $\theta^0 = 1$ and $\pi^0 = \alpha^0 = (\frac{1}{64}, \ldots, \frac{1}{64})$. The stopping criterion for Newton's method was tuned to achieve precision comparable to the stopping criterion (2.6) for the MM algorithms. Running times in seconds were recorded from a laptop computer.

Inspection of Table 3 demonstrates that the MM algorithms outperform Newton's method and that acceleration is often very beneficial. The cost of evaluating and inverting the observed information matrices of the Neerchal–Morel model significantly slows Newton's method even in these problems with only 64 parameters. The observed information matrix of the Dirichlet-multinomial distribution possesses a special structure (diagonal plus rank-1 perturbation) that makes matrix inversion far easier. Table 3 does not show the human effort in devising, programming, and debugging the various algorithms. For Newton's method, derivation and programming took in excess of one day. Formulas for the score and observed information of the Dirichlet-multinomial and Neerchal–Morel distributions are omitted for the sake of brevity. Fisher's scoring algorithm was not implemented because it is even more cumbersome than Newton's method (Neerchal and Morel 2005).

This numerical comparison is merely for illustrative purpose. Numerical analysts have developed quasi-Newton algorithms to mend the defects of Newton's method. The limited-memory BFGS (LBFGS) algorithm (Nocedal and Wright 2006) is especially pertinent to high-dimensional problems. A systematic comparison of the two methods is worth pursuing.

Table 3. Numerical experiment. Row 1: MM; Row 2: SqMPE1 MM; Row 3: SqRRE1 MM; Row 4: Newton's method using the (`fmincon`) function available in the Matlab Optimization Toolbox.

	DM (π, θ)			DM (α)			Neerchal–Morel		
Digit	L	# iters	Time	L	# iters	Time	L	# iters	Time
0	−37,358	232	0.18	−37,358	361	0.16	−38,828	15	0.08
	−37,358	18	0.04	−37,358	18	0.04	−38,828	7	0.10
	−37,358	21	0.04	−37,358	18	0.04	−38,828	7	0.09
	−37,359	11	0.13	−37,358	18	0.16	−38,828	13	106.42
1	−42,179	237	0.16	−42,179	120	0.06	−52,424	17	0.09
	−42,179	17	0.03	−42,179	12	0.03	−52,424	7	0.10
	−42,179	26	0.05	−42,179	13	0.03	−52,424	7	0.10
	−42,179	15	0.19	−42,179	14	0.13	−52,424	12	98.91
2	−39,985	213	0.14	−39,985	136	0.07	−47,723	14	0.07
	−39,985	17	0.04	−39,985	15	0.03	−47,723	6	0.08
	−39,985	17	0.04	−39,985	11	0.03	−47,723	6	0.08
	−39,986	15	0.19	−39,985	15	0.13	−47,721	14	113.15
3	−40,519	214	0.14	−40,519	173	0.08	−45,816	14	0.07
	−40,519	23	0.04	−40,519	15	0.03	−45,816	6	0.08
	−40,519	20	0.04	−40,519	11	0.03	−45,816	6	0.08
	−40,519	14	0.17	−40,519	15	0.13	−45,816	12	102.30
4	−43,489	203	0.13	−43,489	102	0.06	−55,432	14	0.07
	−43,489	17	0.04	−43,489	12	0.03	−55,432	6	0.08
	−43,489	19	0.04	−43,489	9	0.03	−55,432	6	0.08
	−43,489	13	0.17	−43,489	14	0.12	−55,432	14	114.40
5	−41,191	205	0.13	−41,191	116	0.06	−50,063	13	0.07
	−41,191	18	0.04	−41,191	12	0.03	−50,063	6	0.08
	−41,191	19	0.04	−41,191	12	0.03	−50,063	6	0.09
	−41,192	12	0.16	−41,191	15	0.13	−50,063	15	118.22
6	−37,703	232	0.15	−37,703	203	0.10	−41,888	20	0.10
	−37,703	19	0.04	−37,703	16	0.03	−41,888	8	0.11
	−37,703	21	0.04	−37,703	11	0.03	−41,888	8	0.11
	−37,703	15	0.19	−37,703	19	0.16	−41,888	13	104.25
7	−40,304	218	0.14	−40,304	141	0.07	−47,653	12	0.06
	−40,304	16	0.04	−40,304	15	0.03	−47,653	6	0.08
	−40,304	18	0.04	−40,304	11	0.03	−47,653	6	0.08
	−40,305	13	0.15	−40,304	15	0.13	−47,653	15	120.95
8	−43,131	227	0.15	−43,131	171	0.08	−48,844	17	0.09
	−43,131	19	0.04	−43,131	16	0.03	−48,844	7	0.10
	−43,131	23	0.04	−43,131	14	0.03	−48,844	7	0.09
	−43,132	10	0.13	−43,131	15	0.14	−48,844	13	107.22
9	−43,710	207	0.14	−43,710	116	0.06	−53,030	13	0.07
	−43,710	19	0.04	−43,710	12	0.03	−53,030	6	0.08
	−43,710	18	0.04	−43,710	11	0.03	−53,030	6	0.08
	−43,710	12	0.16	−43,710	15	0.14	−53,030	14	116.49

5. DISCUSSION

In designing algorithms for maximum likelihood estimation, Newton's method and Fisher scoring come immediately to mind. In the last generation, statisticians have added

the EM principle. These are good mental reflexes, but the broader MM principle also deserves serious consideration. In many problems, the EM and MM perspectives lead to the same algorithm. In other situations such as image reconstruction in transmission tomography, it is possible to construct different EM and MM algorithms for the same purpose (Lange 2004). One of the most appealing features of the EM perspective is that it provides a statistical interpretation of algorithm intermediates. Although it is a matter of taste and experience whether inequalities or missing data offer an easier path to algorithm development, the fact that there are two routes adds to the possibilities for new algorithms.

One can argue that applications of minorizations (2.3) and (2.5) are just disguised EM algorithms. This objection misses the point in three ways. First, it does not suggest missing data structures explaining the minorization (2.4) and other less well-know minorizations. Second, it fails to weigh the difficulties of invoking simple inequalities versus calculating conditional expectations. When the creation of an appropriate surrogate function requires several minorizations, the corresponding conditional expectations become harder to execute. For example, although the EM principle dictates adding pseudo-observations for zero-truncated data, it is easy to lose sight of this simple interpretation in complicated examples such as the beta-binomial distribution. The genetic segregation analysis example appearing in chapter 2 of the book by Lange (2002) falls into the same category. Third, it fails to acknowledge the conceptual clarity of the MM principle, which shifts focus away from the probability spaces connected with missing data to the simple act of minorization. For instance, when one undertakes maximum a posteriori estimation, should the E step of the EM algorithm take into account the prior?

Some EM and MM algorithms are notoriously slow to converge. As we noted earlier, slow convergence is partially offset by the simplicity of each iteration. There is a growing body of techniques for accelerating MM algorithms (Jamshidian and Jennrich 1995; Lange 1995a; Jamshidian and Jennrich 1997; Varadhan and Rolland 2008). These techniques often lead to a ten-fold or even a hundred-fold reduction in the number of iterations. The various examples appearing in this article are typical in this regard. On problems with boundaries or nondifferentiable objective functions, acceleration may be less helpful.

Our negative-multinomial example highlights two useful tactics for overcoming complications in solving the maximization step of the EM and MM algorithms. It is a mistake to think of the various optimization algorithms in isolation. Often block relaxation (de Leeuw 1994) and Newton's method can be combined creatively with the MM principle. Systematic application of Newton's method in solving the maximization step of the MM algorithm is formalized in the MM gradient algorithm (Lange 1995b).

Parameter asymptotic standard errors are a natural byproduct of Newton's method and scoring. With a modicum of additional effort, the EM and MM algorithms also deliver asymptotic standard errors (Meng and Rubin 1991; Hunter and Lange 2004). Virtually all optimization algorithms are prone to converge to inferior modes. For this reason, we have emphasized finding reasonable initial values. The overlooked article of Ueda and Nakano (1998) suggested an annealing approach to maximization with mixture models. Here the idea is to flatten the likelihood surface and eliminate all but the dominant mode. As the iterations proceed, the flat surface gradually warps into the true bumpy surface. Our recent

work (Zhou and Lange 2010) extends this idea to many other EM and MM algorithms. A similar idea, called graduated non-convexity (GNC), appears in computer vision and signal processing literature (Blake and Zisserman 1987). In the absence of a good annealing procedure, one can fall back on starting an optimization algorithm from multiple random points, but this inevitably increases the computational load. The reassurance that a log-likelihood is concave is always welcome.

Readers may want to try their hands at devising their own MM algorithms. For instance, the Dirichlet-negative-multinomial distribution, the bivariate Poisson (Johnson, Kotz, and Balakrishnan 1997), and truncated multivariate discrete distributions yield readily to the techniques described. The performance of the MM algorithm on these problems is similar to that in our fully developed examples. Of course, many objective functions are very complicated, and devising a good MM algorithm is a challenge. The greatest payoffs are apt to be on high-dimensional problems. For simplicity of exposition, we have not tackled any extremely high-dimensional problems, but these certainly exist (Sabatti and Lange 2002; Ayers and Lange 2008; Lange and Wu 2008). In any event, most mathematicians and statisticians keep a few tricks up their sleeves. The MM principle belongs there, waiting for the right problems to come along.

SUPPLEMENTAL MATERIALS

Datasets and Matlab codes: The supplementary material (a single `zip` package) contains all datasets appearing here and the Matlab codes generating our numerical results and graphs. The `readme.txt` file describes the contents of each file in the package. (supp_material.zip)

ACKNOWLEDGMENTS

The authors thank the editors and referees for their many valuable comments.

[Received February 2009. Revised November 2009.]

REFERENCES

Asuncion, A., and Newman, D. J. (2007), "(UCI) Machine Learning Repository," available at *http://www.ics.uci.edu/~mlearn/Repository.html*. [660]

Ayers, K. L., and Lange, K. (2008), "Penalized Estimation of Haplotype Frequencies," *Bioinformatics*, 24, 1596–1602. [663]

Bailey, N. T. J. (1957), *The Mathematical Theory of Epidemics*, London: Charles Griffin & Company. [649]

Blake, A., and Zisserman, A. (1987), *Visual Reconstruction*, Cambridge, MA: MIT Press. [663]

Bouguila, N. (2008), "Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions," *IEEE Transactions on Knowledge and Data Engineering*, 20 (4), 462–474. [652,660]

Charalambides, C. A., (2007), "Distributions of Random Partitions and Their Applications," *Methodology and Computing in Applied Probability*, 9, 163–193. [658]

Connor, R. J., and Mosimann, J. E. (1969), "Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194–206. [651]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 39 (1), 1–38. [645]

de Leeuw, J. (1994), "Block Relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, eds. H. H. Bock, W. Lenski, and M. M. Richter, Berlin: Springer-Verlag. [645,656,662]

Ewens, W. J. (2004), *Mathematical Population Genetics* (2nd ed.), New York: Springer-Verlag. [657]

Griffiths, D. A. (1973), "Maximum Likelihood Estimation for the Beta-Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease," *Biometrics*, 29, 637–648. [649]

Groenen, P. J. F. (1993), *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*, Leiden, The Netherlands: DSWO Press. [645]

Haldane, J. B. S. (1941), "The Fitting of Binomial Distributions," *Annals of Eugenics*, 11, 179–181. [646,649]

Haseman, J. K., and Soares, E. R. (1976), "The Distribution of Fetal Death in Control Mice and Its Implications on Statistical Tests for Dominant Lethal Effects," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 41, 277–288. [651]

Heiser, W. J. (1995), "Convergent Computing by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis," in *Recent Advances in Descriptive Multivariate Analysis*, ed. W. J. Krzanowski, Oxford: Clarendon Press, pp. 157–189. [645]

Hunter, D. R., and Lange, K. (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37. [645,662]

Jamshidian, M., and Jennrich, R. I. (1995), "Acceleration of the EM Algorithm by Using Quasi-Newton Methods," *Journal of the Royal Statistical Society, Ser. B*, 59, 569–587. [648,662]

——— (1997), "Quasi-Newton Acceleration of the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 59, 569–587. [648,662]

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, New York: Wiley. [657,663]

Lange, K. (1995a), "A Quasi-Newton Acceleration of the EM Algorithm," *Statistica Sinica*, 5, 1–18. [648,662]

——— (1995b), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 57 (2), 425–437. [662]

——— (2002), *Mathematical and Statistical Methods for Genetic Analysis* (2nd ed.), New York: Springer-Verlag. [662]

——— (2004), *Optimization*, New York: Springer-Verlag. [645,647,662]

Lange, K., and Wu, T. T. (2008), "An MM Algorithm for Multicategory Vertex Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 17, 527–544. [663]

Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909. [662]

Minka, T. P. (2003), "Estimating a Dirichlet Distribution," Technical report, Microsoft. [650]

Neerchal, N. K., and Morel, J. G. (1998), "Large Cluster Results for Two Parametric Multinomial Extra Variation Models," *Journal of the American Statistical Association*, 93 (443), 1078–1087. [653]

——— (2005), "An Improved Method for the Computation of Maximum Likelihood Estimates for Multinomial Overdispersion Models," *Computational Statistics & Data Analysis*, 49 (1), 33–43. [653,660]

Nobuaki, H. (2001), "Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment," *Journal of Official Statistics*, 17, 499–520. [658]

Nocedal, J., and Wright, S. (2006), *Numerical Optimization*, New York: Springer. [660]

Patil, M. K., and Shirke, D. T. (2007), "Testing Parameter of the Power Series Distribution of a Zero Inflated Power Series Model," *Statistical Methodology*, 4, 393–406. [660]

Paul, S. R., Balasooriya, U., and Banerjee, T. (2005), "Fisher Information Matrix of the Dirichlet-Multinomial Distribution," *Biometrical Journal*, 47 (2), 230–236. [651]

Pitman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory and Related Fields*, 102 (2), 145–158. [657]

Sabatti, C., and Lange, K. (2002), "Genomewide Motif Identification Using a Dictionary Model," *Proceedings of the IEEE*, 90, 1803–1810. [663]

Steele, J. M. (2004), *The Cauchy–Schwarz Master Class. MAA Problem Books Series*, Washington, DC: Mathematical Association of America. [647]

Ueda, N., and Nakano, R. (1998), "Deterministic Annealing EM Algorithm," *Neural Networks*, 11, 271–282. [662]

Varadhan, R., and Roland, C. (2008), "Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm," *Scandinavian Journal of Statistics*, 35, 335–353. [648,662]

Wu, T. T., and Lange K. (2010), "The MM Alternative to EM," *Statistical Science*, to appear. [645]

Zhou, H., and Lange, K. (2010), "On the Bumpy Road to the Dominant Mode," *Scandinavian Journal of Statistics*, to appear. [663]