Published in final edited form as:
Comput Stat Data Anal. 2012 December ; 56(12): 3909–3920. doi:10.1016/j.csda.2012.05.018.

EM vs MM: A Case Study

Hua Zhou[*] and **Yiwen Zhang**
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA

Abstract

The celebrated expectation-maximization (EM) algorithm is one of the most widely used optimization methods in statistics. In recent years it has been realized that EM algorithm is a special case of the more general minorization-maximization (MM) principle. Both algorithms creates a surrogate function in the first (E or M) step that is maximized in the second M step. This two step process always drives the objective function uphill and is iterated until the parameters converge. The two algorithms differ in the way the surrogate function is constructed. The expectation step of the EM algorithm relies on calculating conditional expectations, while the minorization step of the MM algorithm builds on crafty use of inequalities. For many problems, EM and MM derivations yield the same algorithm. This expository note walks through the construction of both algorithms for estimating the parameters of the Dirichlet-Multinomial distribution. This particular case is of interest because EM and MM derivations lead to two different algorithms with completely distinct operating characteristics. The EM algorithm converges fast but involves solving a nontrivial maximization problem in the M step. In contrast the MM updates are extremely simple but converge slowly. An EM-MM hybrid algorithm is derived which shows faster convergence than the MM algorithm in certain parameter regimes. The local convergence rates of the three algorithms are studied theoretically from the unifying MM point of view and also compared on numerical examples.

Keywords

Convergence rate; Dirichlet-Multinomial distribution; EM algorithm; MM algorithm

1. Introduction

Numerical optimization methods have been intensively used by statisticians due to the popularity of the maximum likelihood estimation. A powerful weapon among them is the celebrated expectation-maximization (EM) algorithm (Dempster et al., 1977). The E step in the EM algorithm creates a Q function which is then minimized in the M step. In many problems the surrogate Q function is much simpler than the log-likelihood and thus the M step can be solved analytically. These two steps iterate until the parameters converge. In recent years, it has been realized that the EM algorithm is a special case of the more general minorization-maximization (MM) principle (de Leeuw, 1994; Heiser, 1995; Lange et al., 2000; Wu and Lange, 2010). The first M step of an MM algorithm creates a minorizing function that is optimized in the second M step. The Q function in the EM algorithm is a

[*]Corresponding author. Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA, Phone: 1-919-515-2570, Fax: 1-919-515-7591. hua_zhou@ncsu.edu (Hua Zhou).

specific example of minorizing functions. Same as the EM algorithm, this two-step process always drives the objective function uphill. The key difference between the two algorithms is the construction of surrogate functions. The minorization step in the MM algorithm hinges upon recognizing and manipulating inequalities, while EM algorithm relies on calculating conditional expectations. Advantages enjoyed by both algorithms are their numerical stability, natural adaption to parameter constraints, and scalability to high dimensions. An open question has been raised whether any MM algorithm can be recast as an EM algorithm (Meng, 2000). For instance, the MM algorithms for fitting Bradley-Terry models (Hunter, 2004) have recently been shown to be equivalent to EM algorithms with appropriately chosen latent variables (Caron and Doucet, 2010). However, in these worked out examples, the construction of missing data framework turns out non-intuitive and irrelevant to the statistical model that generates the data. Taking the MM point of view frees the derivation from the dependence on a missing data framework. For instance, the recent article (Wu and Lange, 2010) demonstrates the potential of the MM algorithm in random graph models, discriminant analysis and image restoration problems where there is no apparent missing data structure.

This expository paper walks through the construction of both EM and MM algorithms for the maximum likelihood estimation of the Dirichlet-Multinomial distribution. For this particular problem they produce two completely different algorithms. The Q function in the EM algorithm is fraught with special functions (digamma and trigamma) and the M step resists analytical solutions and has to resort to iterative, multivariate Newton's method. In contrast, the surrogate function of the MM algorithm is much simpler and yields trivial updates in the M step. Re-inspecting the M step of EM algorithm from the MM perspective leads to an EM-MM hybrid algorithm which partially resolves the difficulty in the M step of the EM algorithm. Similar hybrid algorithm is utilized in fitting mixture of Plackett-Luce models for ranking data (Gromley and Murphy, 2008). The local convergence rates of the MM and hybrid algorithms are studied theoretically and demonstrated on numerical experiments. There is no clear winner in the sense that one converges faster than the others in all parameter regimes.

As a road map to the remainder of the paper, Section 2 lays out the problem being studied and introduces a classical data set for numerical illustrations. EM and MM algorithms are derived in Section 3 and 4 respectively. The difficulty in maximizing the Q function in the EM algorithm can be partially remedied if we take the MM point of view. This connection is explored in Section 5 and leads to an EM-MM hybrid algorithm. Local convergence properties of the three algorithms are studied in Section 6. The operating characteristics (run time, convergence rates, and final objective values) of the three algorithms are compared numerically under various parameter settings in Section 7. Finally we conclude with some other options for solving this problem.

2. Problem Setup and a Running Example

Multivariate count data frequently arise in genetics (Lange, 2002; Tvedebrink, 2010; Ionita-Laza and Laird, 2010), toxicology (Hines and Lawless, 1993), protein homology detection (Sjölander et al., 1996), word burstiness modeling (Madsen et al., 2005), and language modeling (MacKay and Bauman Peto, 1994). When multivariate count data exhibit over-dispersion, the Dirichlet-Multinomial distribution is preferred over the familiar multinomial distribution. In the Dirichlet-Multinomial sampling, the multinomial parameter $\boldsymbol{p} = (p_1,\ldots, p_d)$ is modeled as a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1,\ldots, \alpha_d)$, where $\alpha_j > 0$.

Accordingly, given a multivariate count vector $\boldsymbol{x} = (x_1, \ldots, x_d)$ with batch size $m = \sum_{j=1}^{d} x_j$, the probability mass function under a Dirichlet-Multinomial model is

$$\begin{aligned} f(\mathrm{x}|\alpha) &= \int_{\Delta_d} \binom{m}{\mathrm{x}} \prod_{j=1}^{d} p_j^{x_j} \frac{\Gamma(|\alpha|)}{\prod_{j=1}^{d} \Gamma(\alpha_j)} \prod_{j=1}^{d} p_j^{\alpha_j - 1} d\boldsymbol{p} \\ &= \binom{m}{\boldsymbol{x}} \frac{\prod_{j=1}^{d} \Gamma(\alpha_j + x_j)}{\Gamma(|\alpha| + m)} \frac{\Gamma(|\alpha|)}{\prod_{j=1}^{d} \Gamma(\alpha_j)} \end{aligned} \quad (1)$$

$$= \binom{m}{\boldsymbol{x}} \frac{\prod_{j=1}^{d} (\alpha_j)_{x_j}}{|\alpha|_m}, \quad (2)$$

where $|\alpha| = \sum_{j=1}^{d} \alpha_j$ and $(a)_k = \prod_{i=0}^{k-1} (a+i)$ denotes the rising factorial. The last equality is due to the fact $\Gamma(a + k)/\Gamma(a) = (a)_k$. An alternative parametrization uses

$$\pi_j = \frac{\alpha_j}{|\alpha|}, \quad j = 1, \ldots, d, \quad \theta = \frac{1}{|\alpha|}, \quad (3)$$

in terms of the proportion vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$ and the over-dispersion parameter θ. For the sake of brevity, we stick to parametrization (2) in this article. The derivation and most conclusions equally apply to both parameterizations. Given independent data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the log-likelihood is

$$\begin{aligned} l(\alpha) &= \sum_{i=1}^{n} \ln f(\boldsymbol{x}_i|\alpha) \\ &= \sum_{i=1}^{n} \ln \binom{m_i}{\boldsymbol{x}_i} + \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=0}^{x_{ij}-1} \ln(\alpha_j + k) - \sum_{i=1}^{n} \sum_{k=0}^{m_i - 1} \ln(|\alpha| + k) \end{aligned} \quad (4)$$

and the maximum likelihood estimation seeks the maximizer of (4). Most current applications utilize the Newton's method for finding the MLE (Lange, 2002; Tvedebrink, 2010; Ionita-Laza and Laird, 2010), which may be numerically instable because the objective function (4) is non-concave. The alternative Fisher's scoring algorithm replaces the observed information matrix in Newton's method by expected information matrix and yields an ascent algorithm. However, the calculation of expected information matrix for Dirichlet-Multinomial model is expensive due to numerous evaluations of beta-binomial tail probabilities (Paul et al., 2005). Recently Zhou and Lange (2010) devise the MM algorithm for a whole class of multivariate discrete distributions which include the Dirichlet-Multinomial as a special case. Compared to the Newton's method, the MM algorithm is numerically stable, easy to implement and scalable to high-dimensional data. In this article, we take up the alternative EM approach for maximizing the log-likelihood (4) and contrast it to the MM algorithm in respect to algorithmic design, per iteration computation cost, and local convergence rate.

As a numerical example we consider the classical data on the mice that are exposed to various mutagens (Haseman and Soares, 1976). Environmental scientists are interested in investigating the mutagenicity of a compound or irradiation in vivo in mice. Male mice are treated with the suspect mutagen and then paired to one or more female mice. Seventeen days after the initial exposure to a male, females are killed and their uteri are examined for the presence of living and dead embryos (implants). In the first data set of Haseman and Soares (1976), denoted by HS76-1 in following, there are $n = 524$ females with the total

number of implants per female m_i varying from 1 to 20. Counts of dead and survived implants are recorded for each female. Figure 2 displays the histogram of the 524 proportions of dead implants. The variability in the proportions is prominent and the traditional binomial distribution is inappropriate for such over-dispersion count data. Fitting the beta-binomial distribution ($d = 2$) to the HS76-1 data set gives the MLE $\hat{\boldsymbol{\alpha}} = (1.23, 12.46)$ with log-likelihood −777.79. The density of the beta distribution with parameter $\hat{\boldsymbol{\alpha}}$ is imposed on the histogram in Figure 2 and demonstrates a good fit. The classical binomial fit gives a log-likelihood −842.61. The likelihood ratio test of the over-dispersion parameter $H_0 : \theta = (\alpha_1 + \alpha_2)^{-1} = 0$ vs $H_1 : \theta > 0$ yields a p-value essentially 0, corroborating the appropriateness of beta-binomial model.

The performances of the EM, MM and a hybrid algorithm on this data set will be compared in Section 6. Although we use a $d = 2$ data set as the running example for ease of illustration and visualization, all our derivations and convergence rate results are for general d and numerical experiments are carried out for d as high as 50 in Section 7. We begin with the EM algorithm for maximizing (4).

3. EM Algorithm

Derivation of EM algorithm hinges upon a missing data structure. Let $f(\boldsymbol{\theta})$ be the log-likelihood of the observed data with parameter vector $\boldsymbol{\theta}$. In the E step, a surrogate function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is calculated as the conditional expectation of the complete data log-likelihood given current parameter iterate $\boldsymbol{\theta}^{(t)}$. The well-known calculations (Baum et al., 1970; Dempster et al., 1977) demonstrate that the Q function satisfies the fundamental inequality

$$f(\theta) - f(\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}). \quad (5)$$

Maximizing the surrogate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ generates the next iterate $\boldsymbol{\theta}^{(t+1)}$ which obviously drives the log-likelihood of the observed data uphill.

The admixture representation (1) of the Dirichlet-Multinomial distribution naturally implies an EM algorithm. We consider $\boldsymbol{p}_i$, $i = 1, \ldots, n$, as the missing data and denote the joint density of complete data by $\prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{p}_i, \alpha)$. Then the Q function is

$$Q(\alpha|\alpha^{(t)}) = \mathbf{E}[\ln \prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{p}_i, \alpha)|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \alpha^{(t)}],$$

where the expectation is with respect to the conditional distribution

$$f(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \alpha^{(t)}),$$

i.e., independent Dirichlet($\boldsymbol{x}_i + \boldsymbol{\alpha}^{(t)}$). Therefore

$$Q(\alpha|\alpha^{(t)}) = \sum_{i=1}^{n} Q_i(\alpha|\alpha^{(t)})$$

with

$$
\begin{aligned}
& Q_i(\alpha|\alpha^{(t)}) \\
= & \int_{\Delta_d} \frac{\Gamma(m_i+|\alpha^{(t)}|)}{\prod_{j=1}^d \Gamma(x_{ij}+\alpha_j^{(t)})} \prod_j p_{ij}^{x_{ij}+\alpha_j^{(t)}-1} \\
& \cdot \left[\ln \binom{m_i}{\boldsymbol{x}_i} + \sum_{j=1}^d (x_{ij}+\alpha_j - 1) \ln p_{ij} + \ln \Gamma(|\alpha|) - \sum_{j=1}^d \ln \Gamma(\alpha_j) \right] d\boldsymbol{p}_i \\
= & \ln \binom{m_i}{\boldsymbol{x}_i} + \sum_{j=1}^d (x_{ij}+\alpha_j - 1) \left[\Psi(x_{ij}+\alpha_j^{(t)}) - \Psi(m_i+|\alpha^{(t)}|) \right] \\
& + \ln \Gamma(|\alpha|) - \sum_{j=1}^d \ln \Gamma(\alpha_j).
\end{aligned}
$$

Here $\Psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function and the exponential family differential identity is used to calculate the expectation $\mathbf{E}[\ln P_j]$ under a Dirichlet distribution. Therefore

$$
\begin{aligned}
Q(\alpha|\alpha^{(t)}) &= \sum_{i=1}^n \ln \binom{m_i}{\boldsymbol{x}_i} + \sum_{i=1}^n \sum_{j=1}^d (x_{ij}+\alpha_j - 1) \left[\Psi(x_{ij}+\alpha_j^{(t)}) - \Psi(m_i+|\alpha^{(t)}|) \right] \\
&\quad + n \ln \Gamma(|\alpha|) - n \sum_{j=1}^d \ln \Gamma(\alpha_j) \\
&= \sum_{j=1}^d \sum_{i=1}^n \alpha_j \left[\Psi(x_{ij}+\alpha_j^{(t)}) - \Psi(m_i+|\alpha^{(t)}|) \right] - n \sum_{j=1}^d \ln \Gamma(\alpha_j) \\
&\quad + n \ln \Gamma(|\alpha|) + c^{(t)}.
\end{aligned} \tag{6}
$$

Throughout the paper we use $c^{(t)}$ to collect constants that are irrelevant to the optimization and it may vary in different equations. Maximizing $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)})$ is not trivial since α_j are intertwined in the $\ln \Gamma(|\boldsymbol{\alpha}|)$ term. The Newton's method has to be utilized for the M step. The first two derivatives of $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)})$ are

$$
\begin{aligned}
[\nabla Q(\alpha|\alpha^{(t)})]_j &= \frac{\partial Q}{\partial \alpha_j} = \sum_i [\Psi(x_{ij}+\alpha_j^{(t)}) - \Psi(m_i+|\alpha^{(t)}|)] + n\Psi(|\alpha|) - n\Psi(\alpha_j) \\
[d^2 Q(\alpha|\alpha^{(t)})]_{jj'} &= \frac{\partial^2 Q}{\partial \alpha_j \, \alpha_{j'}} = n\psi(|\alpha|) - n\psi(\alpha_j) 1_{\{j=j'\}},
\end{aligned}
$$

where $\psi(z) = \Psi'(z)$ is the trigamma function. Newton's method iterates according to

$$
\alpha_{(m+1)} = \alpha_{(m)} - \left[d^2 Q(\alpha_{(m)}|\alpha^{(t)}) \right]^{-1} \cdot \nabla Q(\alpha_{(m)}|\alpha^{(t)}),
$$

where the subscript m indicates its iteration number. Several issues arise here. First, in each iteration the Hessian matrix $d^2 Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(t)})$ has to be computed and a linear system needs to be solved. Second, since the Q function is non-concave (note $\ln \Gamma$ is convex), the Newton's method may not generate an ascent algorithm. Even when its Hessian is negative definite locally, a line search strategy may be necessary to prevent over-shooting. Lastly the Newton's updates may violate the parameter constraints $\alpha_j > 0$. At this point it is realized that the EM principle has not reduced the difficulty of the original optimization problem and the Newton's or Fisher's scoring method could be used directly on the observed data log-likelihood (4). However there is a remedy. Before that we first explore the MM solution.

4. MM Algorithm

Like EM, the MM algorithm is a general principle for creating optimization algorithms. The survey papers (Lange et al., 2000; Hunter and Lange, 2004) and textbook treatment (Lange, 2010) serve as an excellent introduction. The derivation in this section also appears in (Zhou and Lange, 2010) as a special case. Let $f(\boldsymbol{\theta})$ be the objective function, not necessarily a loglikelihood, whose maximum we seek. An MM algorithm involves minorizing $f(\boldsymbol{\theta})$ at current iterate $\boldsymbol{\theta}^{(t)}$ by a surrogate function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ that satisfies two properties

$$\begin{aligned} f(\theta) &\geq g(\theta|\theta^{(t)}), \qquad \theta \neq \theta^{(t)} \\ f(\theta^{(t)}) &= g(\theta^{(t)}|\theta^{(t)}). \end{aligned}$$

In other words, the surface $\boldsymbol{\theta} \mapsto g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ lies below the surface $\boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ and is tangent to it at the current iteration $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. The construction of the minorizing function $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ constitutes the first M of the MM algorithm. The second M of the MM algorithm maximizes the surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than $f(\boldsymbol{\theta})$ directly. If $\boldsymbol{\theta}^{(t+1)}$ denotes the maximum of $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to its left argument, then $\boldsymbol{\theta}^{(t+1)}$ increases $f(\boldsymbol{\theta})$. It follows directly from the inequalities

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

This ascent property is the source of the MM algorithm's numerical stability and remains valid if we merely increase $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ rather than maximize it.

The fundamental inequality (5) in the EM algorithm shows that the Q function produced in the E step constitutes a minorizing function of the log-likelihood up to an additive constant. This fact readily qualifies EM algorithm as a special case of the MM algorithm. The MM perspective is more general as it frees algorithm derivation from the missing data straitjacket and invites wider applications. Wu and Lange (2010) briefly summarize the history of the MM algorithm and showcase its flexibility in some problems in which the EM derivation is hard to carry out.

To construct an MM algorithm for maximizing the Dirichlet-Multinomial log-likelihood function (4), the strategy is to minorize term by term. We first simplify the two sums

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=0}^{x_{ij}-1} \ln(\alpha_j + k) &= \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=0}^{\max_i x_{ij}-1} \ln(\alpha_j + k) 1_{\{x_{ij}-1 \geq k\}} \\ &= \sum_{j=1}^{d} \sum_{k=0}^{\max_i x_{ij}-1} \ln(\alpha_j + k) \sum_{i=1}^{n} 1_{\{k \leq x_{ij}-1\}} = \sum_{j=1}^{d} \sum_{k=0}^{\max_i x_{ij}-1} s_{jk} \ln(\alpha_j + k) \end{aligned}$$

and

$$\begin{aligned} -\sum_{i=0}^{n} \sum_{k=0}^{m_i-1} \ln(|\alpha| + k) &= -\sum_{k=0}^{m_i-1} \ln(|\alpha| + k) \sum_{i=1}^{n} 1_{\{k \leq m_i-1\}} \\ &= -\sum_{k=0}^{\max_i m_i-1} r_k \ln(|\alpha| + k), \end{aligned}$$

where

$$s_{jk}=\sum_{i=1}^{n}1_{\{x_{ij}\geq k+1\}},\quad r_k=\sum_{i=1}^{n}1_{m_i\geq k+1}$$

are counts. Applying the Jensen's inequality to the $\ln(\alpha_j + k)$ terms

$$\begin{aligned}\ln(\alpha_j+k) &\geq \frac{\alpha_j^{(t)}}{\alpha_j^{(t)}+k}\ln\left(\frac{\alpha_j^{(t)}+k}{\alpha_j^{(t)}}\cdot\alpha_j\right)+\frac{k}{\alpha_j^{(t)}+k}\ln\left(\frac{\alpha_j^{(t)}+k}{k}\cdot k\right)\\ &= \frac{\alpha_j^{(t)}}{\alpha_j^{(t)}+k}\ln\alpha_j+c^{(t)}\end{aligned}\tag{7}$$

and the supporting hyperplane inequality to the $-\ln(|\boldsymbol{\alpha}| + k)$ terms

$$\begin{aligned}-\ln(|\alpha|+k) &\geq -\frac{|\alpha|-|\alpha^{(t)}|}{|\alpha^{(t)}|+k}-\ln(|\alpha^{(t)}|+k)\\ &= -\frac{|\alpha|}{|\alpha^{(t)}|+k}+c^{(t)}\end{aligned}\tag{8}$$

yields the surrogate function

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

Figure 2 depicts the two minorization inequalities (7) and (8) with $\boldsymbol{\alpha}^{(t)} = 2$ and $k = 1$. In both minorizations, the equality holds at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(t)}$. Finding the $\boldsymbol{\alpha}$ that maximizes $g(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{(t)})$ is trivial and leads to the multiplicative MM updates

$$\alpha_j^{(t+1)}=\alpha_j^{(t)}\frac{\sum_k\frac{s_{jk}}{\alpha_j^{(t)}+k}}{\sum_k\frac{r_k}{|\alpha^{(t)}|+k}},\quad j=1,\ldots,d,\tag{9}$$

which are substantially simpler than their EM counterparts. It is noteworthy that the parameter constraints $\alpha_j > 0$ are always satisfied whenever $\alpha_j^{(0)}>0$. Besides offering a completely different algorithm from EM, MM principle also suggests a remedy for the troublesome M step in the EM algorithm.

5. An EM-MM Hybrid Algorithm

Due to the fact that the $\ln\Gamma(x)$ function is convex, we can resort to the supporting hyperplane inequality to separate the parameters in the $\ln\Gamma(|\boldsymbol{\alpha}|)$ term in the Q function (6)

$$\begin{aligned}&Q(\alpha|\alpha^{(t)})\\ \geq\ &\sum_i\ln\binom{m_i}{\boldsymbol{x}_i}+\sum_i\sum_j(x_{ij}+\alpha_j-1)\left[\Psi(x_{ij}+\alpha_j^{(t)})-\Psi(m_i+|\alpha^{(t)}|)\right]\\ &+n\Psi(|\alpha^{(t)}|)(|\alpha|-|\alpha^{(t)}|)+n\ln\Gamma(|\alpha^{(t)}|)-n\sum_j\ln\Gamma(\alpha_j)+c^{(t)}\\ =\ &\sum_j\left\{\alpha_j\sum_i\left[\Psi(x_{ij}+\alpha_j^{(t)})-\Psi(m_i+|\alpha^{(t)}|)+\Psi(|\alpha^{(t)}|)\right]-n\ln\Gamma(\alpha_j)\right\}+c^{(t)}.\end{aligned}\tag{10}$$

In this new minorizing function (10), the parameters α_j are separated and only need to be optimized independently. Equating the partial derivatives with respect to α_j to 0 gives the function to solve for α_j

$$\sum_i \Psi(x_{ij}+\alpha_j^{(t)}) - n\Psi(\alpha_j) = \sum_i \Psi(m_i + |\alpha^{(t)}|) - n\Psi(|\alpha^{(t)}|).$$

In view of the recurrence relation $\Psi(y+1) = \Psi(y) + 1/y$, this is equivalent to

$$\begin{aligned} \Psi(\alpha_j^{(t)}) - \Psi(\alpha_j) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{m_i-1}\frac{1}{|\alpha^{(t)}|+k} - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{x_{ij}-1}\frac{1}{\alpha_j^{(t)}+k} \\ &= \frac{1}{n}\sum_{k=0}^{\max_i m_i-1}\frac{r_k}{|\alpha^{(t)}|+k} - \frac{1}{n}\sum_{k=0}^{\max_i x_{ij}-1}\frac{s_{jk}}{\alpha_j^{(t)}+k}. \end{aligned}$$

It is interesting to note that the two sums on the right hand side were used in the MM updates (9) in a completely different way. Here we can find the root of $\Psi(\alpha_j) = a$ by Newton's iterates

$$\alpha_{(m+1)} = \alpha_{(m)} - \frac{\Psi(\alpha_{(m)}) - a}{\psi(\alpha_{(m)})}.$$

Since ln Γ is convex, the trigamma function ψ is positive and the Newton's method is guaranteed to converge to the right root. The Newton's method applied to individual α_j is substantially simpler than the multivariate Newton's method in the original EM algorithm, which involves computing and inverting Hessian matrix in each iteration. As we will see in the next section, the price it pays is more iterations.

6. Convergence Rates

The EM, MM, and hybrid algorithms constructed so far enjoy the same ascent property, yet with distinct per iteration computational cost. A formal comparison entails a close study of their local convergence rates, which roughly measure how fast they converge near the optimal point. Figure 3 displays the iterates of the three algorithms for the HS76-1 data set with two different starting points. When starting from a blind guess $\boldsymbol{\alpha}^{(0)} = (1, 1)$, the EM algorithm converges quickly within 20 iterations, while the other two lag behind. The hybrid algorithm outruns MM initially but is caught up when close to the optimal point. When starting from the method of moment estimate $\boldsymbol{\alpha}^{(0)} = (0.4711, 4.8072)$, EM behaves similarly while the MM algorithm narrowly edges out the hybrid algorithm.

Because all three algorithms can be deemed as special cases of the MM principle, their convergence properties can be studied under a unified framework. Consider an MM map $M(\boldsymbol{\theta})$ for maximizing the objective function $f(\boldsymbol{\theta})$ via the surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. When close to the optimal point $\boldsymbol{\theta}^{\infty}$,

$$\theta^{(t+1)} - \theta^{\infty} \approx dM(\theta^{\infty}) \cdot (\theta^{(t)} - \theta^{\infty}),$$

where $dM(\boldsymbol{\theta}^{\infty})$ is the differential of the mapping M at the optimal point $\boldsymbol{\theta}^{\infty}$ of $f(\boldsymbol{\theta})$. Figure 4 displays the distance of the algorithmic iterates to the MLE for the HS76-1 example on the logarithmic scale and shows the linear convergence behavior when close to $\boldsymbol{\theta}^{\infty}$. Therefore

the local convergence rate of the sequence $\boldsymbol{\theta}^{(t+1)} = M(\boldsymbol{\theta}^{(t)})$ is defined as the spectral radius of $dM(\boldsymbol{\theta}^{\infty})$. The familiar calculations (McLachlan and Krishnan, 2008; Lange, 2010) demonstrate that

$$dM(\theta^{\infty}) = \boldsymbol{I} - \left[d^2 g(\theta^{\infty}|\theta^{\infty}) \right]^{-1} d^2 f(\theta^{\infty}). \quad (11)$$

In other words, the local convergence rate is determined by how well the surrogate function g approximates the log-likelihood surface f near the optimal point $\boldsymbol{\theta}^{\infty}$. In EM literature, $dM(\boldsymbol{\theta}^{\infty})$ is called the rate matrix (Meng and Rubin, 1991). A smaller rate means that the surrogate function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{\infty})$ hugs $f(\boldsymbol{\theta})$ tighter around $\boldsymbol{\theta}^{\infty}$ and thus implies faster convergence.

In our case, the Q function of the EM algorithm by construction dominates the minorizing function (10) of the hybrid algorithm. This implies the faster convergence of the EM algorithm than the hybrid algorithm as observed in Figure 3. However, there is no dominance relations between them and the MM surrogate function. Figure 5 displays the log-likelihood surface of the HS76-1 data set and the minorizing functions of the three algorithms at the parameter point $(\alpha_1, \alpha_2) = (0.5, 5)$. All three minorizing functions lie below the log-likelihood surface while touching it at (0.5,5). It is clear that the Q function of the EM algorithm approximates the log-likelihood function better than the minorizing function of the hybrid algorithm over the whole region. The MM minorizing function intersects with the other two besides the point (0.5, 5).

Given a data set, the local convergence rates of the three algorithms can be calculated. Let $\boldsymbol{\alpha}^{\infty}$ be the MLE and define constants

$$\begin{aligned}
a &= \sum_{k=0}^{\max_i m_i - 1} \frac{r_k}{(|\alpha^{\infty}| + k)^2} \\
b_j &= \sum_{k=0}^{\max_i x_{ij} - 1} \frac{s_{jk}}{(\alpha_j^{\infty} + k)^2}, \ j = 1, \ldots, d \\
c_j &= \sum_{k=0}^{\max_i x_{ij} - 1} \frac{s_{jk}}{\alpha_j^{\infty}(\alpha_j^{\infty} + k)}, \ j = 1, \ldots, d \\
d_j &= n\psi(\alpha_j^{\infty}) = \sum_{k=0}^{\infty} \frac{n}{(\alpha_j^{\infty} + k)^2}, \ j = 1, \ldots, d \\
e &= n^{-1}[\psi(|\alpha^{\infty}|)^{-1} - \sum_j d_j^{-1}]
\end{aligned}$$

and three polynomials

$$\begin{aligned}
P_{\mathrm{MM}}(\lambda) &= \prod_{j=1}^{d} (\lambda - b_j c_j^{-1}) + a \sum_{j=1}^{d} c_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} c_{j'}^{-1}) \\
P_{\mathrm{hybrid}}(\lambda) &= \prod_{j=1}^{d} (\lambda - b_j d_j^{-1}) + a \sum_{j=1}^{d} d_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} d_{j'}^{-1}) \\
P_{\mathrm{EM}}(\lambda) &= \prod_{j=1}^{d} (\lambda - b_j d_j^{-1}) - \sum_j (b_j d_j^{-1} e - ae \sum_{j''} d_{j''}^{-1} - a) d_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} d_{j'}^{-1}).
\end{aligned} \quad (12)$$

All roots of these polynomials are real and especially the smallest roots give the information about the local convergence rates of the algorithms. Formally we have the following result.

Proposition 6.1

The MM, hybrid, and EM algorithms have linear rates of convergence $1 - \lambda_{MM}$, $1 - \lambda_{hybrid}$, *and* $1 - \lambda_{EM}$, *respectively, where* λ *are the smallest roots of the corresponding polynomials*.

The proof of Proposition 6.1 utilizes (11) straightforwardly and is relegated to the Appendix. For the HS76-1 data, the local convergence rates of EM, MM, and hybrid algorithms are 0.9893, 0.9915, and 0.9946 respectively, corroborating what we observe in Figure 4. It also illustrates the point that local convergence rate only captures the convergence behavior close to the optimal point. In the left panel of Figure 3, all algorithms are started far away from the optimal point and the hybrid algorithm converges faster than the MM algorithm initially even though it has a slower local convergence rate.

The convergence rates are data dependent and vary with data sets even when they are generated from the same distribution. Approximate convergence rates at any MLE $\boldsymbol{\alpha}^{\infty}$ can be obtained by replacing s_{jk} appearing in (12) by their expected values $\mathbf{E}(s_{jk}) = \sum_{i=1}^{n} \mathbf{P}(X_{ij} > k)$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})$ is a Dirichlet-Multinomial random vector with parameter $\boldsymbol{\alpha}^{\infty}$ and batch size m_i. For the beta-binomial distribution ($d = 2$), the relevant polynomials (12) are quadratic and the smallest roots can be readily computed. Figure 6 displays the approximate convergence rates of the hybrid and MM algorithms over the parameter region $[0.3, 5] \times [0.3, 5]$ for data sets with fixed batch sizes $m = 5$, 10, or 20. The number of observations n cancels in the formula (12) and does not play a role in this setting. The convergence rate of the EM algorithm is not displayed here since it has little practical use. Both algorithms claim their own territory of faster convergence. Neither of them dominates the other in all three settings of m. At $m = 5$, the MM algorithm enjoys faster convergence over the whole region while the reverse is true at $m = 20$. The $m = 10$ case represents a compromise between the two. Both convergence rates improve with increasing batch size m. The convergence is fast for both when α_1 and α_2 are small and slow when their magnitude is large. At fixed $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2$, the convergence is slow when α_j differ dramatically. For general $d > 2$, it is hard to obtain explicit expressions for these rates but the approximate rates can be easily computed and similar observations are made.

7. Numerical Experiments

We stress that the sheer number of iterations until convergence is not the sole determinant of algorithm speed. Computational complexity per iteration also comes into play. The iterations within iteration feature of the EM and hybrid algorithms may compromise their fast convergence in ceratin parameter regimes. To demonstrate the tradeoff between number of iterations and speed of each iteration, we compare the performance of MM, EM, and hybrid algorithms under various parameter values. Figure 7 displays the boxplots of the timing, number of iterations, and final objective values from the three algorithms on 100 multivariate count data sets simulated with parameter values $\boldsymbol{\alpha} = (0.1, 1)$ (row 1), $\boldsymbol{\alpha} = (0.2, 2)$ (row 2), $\alpha_1 = \cdots = \alpha_{50} = 0.5$ (row 3) and $\alpha_1 = \cdots = \alpha_{50} = 5$ (row 4) respectively. The sample size is fixed at $n = 100$ and the batch size at $m = 20$. The trio of algorithms are run on the same data set in each simulation replicate. Convergence is declared when the relative change in objective values is less than 10^{-6}. Apparently there is no winner that can dominate others across all scenarios. The EM algorithm always enjoys the fastest convergence. However its speed is compromised by the extra computation cost per iteration. Convergence rates of MM and hybrid algorithms vary over parameter regime as we already see in Section 6. For high dimensional problems ($d = 50$), the simplicity of MM updates overcomes its slower convergence, making it the fastest one under the two tested parameter settings (rows 3 and 4).

8. Conclusions

Multiple solutions to the same problem have proven illuminating in many areas of mathematics. In this article we devise and compare the EM and MM algorithms for estimating the parameters of Dirichlet-Multinomial distribution. This exercise vividly contrasts the different approaches utilized by EM and MM algorithms for constructing the surrogate functions. In this example, taking the MM perspective remedies the difficulty encountered in the EM algorithm and yields a new algorithm that enjoys faster convergence over certain parameter regimes. This interplay between the two algorithms illustrates the delicacy of algorithmic development in computational statistics.

We have omitted the discussion of two other popular optimization methods, the Newton method and the Fisher's scoring method. The paramount advantage of these two methods is their fast convergence. However they require evaluating and inverting an information matrix at each iteration. For Dirichlet-Multinomial distribution, the information matrices are easy to invert and therefore not a big concern. Even so, the Newton's method may still suffer from instability and violation of parameter constraints. Scoring method remedies the instability by using the expected information matrix, which is negative definite and guarantees an ascent direction. Parameter constraint violation is still pertinent. More severely, calculation of the expected information matrix involves evaluating numerous beta-binomial tail probabilities (Paul et al., 2005). On large scale problems, this is simply infeasible. However there is still room to combine Newton's method and EM or MM algorithms. For example, the recent work (Zhou et al., 2011) presents a new quasi-Newton scheme that achieves one to three orders of magnitude acceleration of many EM and MM algorithms. Many practitioners tend to prefer one particular optimization method over the others. In practice, some of the best algorithms are hybrids.

Acknowledgments

The first author was partially supported by NIH grant R01 HG006139 and NCSU FRPD grant.

Appendix

Proof of Proposition 6.1

The Hessian of the log-likelihood function (4) at the optimal point $\boldsymbol{\alpha}^{\infty}$ is

$$\begin{aligned}[d^2 l(\alpha^{\infty})]_{jj'} &= \sum_k \frac{r_k}{(|\alpha^{\infty}|+k)^2} - \sum_k \frac{s_{jk}}{(\alpha_j^{\infty}+k)^2} 1_{\{j=j'\}} \\ &= a - b_j 1_{\{j=j'\}}.\end{aligned}$$

The MM minorizing function has curvature

$$[d^2 g_{\mathrm{MM}}(\alpha^{\infty}|\alpha^{\infty})]^{-1}_{jj'} = -\left[\sum_k \frac{s_{jk}}{\alpha_j^{\infty}(\alpha_j^{\infty}+k)}\right]^{-1} 1_{\{j=j'\}} = -c_j^{-1} 1_{\{j=j'\}}$$

while the curvature of the minorizing function for the hybrid algorithm is

$$[d^2 g_{\mathrm{hybrid}}(\alpha^{\infty}|\alpha^{\infty})]^{-1}_{jj'} = -[n\psi(\alpha_j^{\infty})]^{-1} 1_{\{j=j'\}} = -d_j^{-1} 1_{\{j=j'\}}.$$

Lastly the EM Q function has curvature

$$\begin{aligned}[d^2 Q(\alpha^\infty|\alpha^\infty)]^{-1}_{jj'} &= -\frac{\psi(\alpha_j^\infty)^{-1}\psi(\alpha_{j'}^\infty)^{-1}}{n[\psi(|\alpha^\infty|)^{-1}-\sum_j \psi(\alpha_j^\infty)^{-1}]} - \frac{\psi(\alpha_j^\infty)^{-1}}{n} 1_{\{j=j'\}} \\ &= -d_j^{-1} d_{j'}^{-1} e^{-1} - d_j^{-1} 1_{\{j=j'\}}.\end{aligned}$$

The characteristic polynomial of $[d^2 g_{\mathrm{MM}}(\boldsymbol{\alpha}^\infty|\boldsymbol{\alpha}^\infty)]^{-1} d^2 f(\boldsymbol{\alpha}^\infty)$ is

$$\begin{aligned}P_{\mathrm{MM}}(\lambda) &= \det[\lambda \boldsymbol{I} - \mathrm{diag}\left(b_j c_j^{-1}\right) + a \cdot \boldsymbol{c}^{-1} 1^t] \\ &= \det[\lambda \boldsymbol{I} - \mathrm{diag}(b_j c_j^{-1})]\{1 + a 1^t [\lambda \boldsymbol{I} - \mathrm{diag}(b_j c_j^{-1})]^{-1} \boldsymbol{c}^{-1}\} \\ &= \prod_{j=1}^{d} (\lambda - b_j c_j^{-1}) + a \sum_{j=1}^{d} c_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} c_{j'}^{-1}).\end{aligned}$$

Here, $\boldsymbol{c}^{-1} = (c_1^{-1}, c_2^{-1}, \ldots, c_d^{-1})'$. Similarly, the characteristic polynomial for the hybrid algorithm is

$$P_{\mathrm{hybrid}}(\lambda) = \prod_{j=1}^{d} (\lambda - b_j d_j^{-1}) + a \sum_{j=1}^{d} d_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} d_{j'}^{-1})$$

and the characteristic polynomial for the EM algorithm is

$$P_{\mathrm{EM}}(\lambda) = \prod_{j=1}^{d} (\lambda - b_j d_j^{-1}) - \sum_j (b_j d_j^{-1} e^{-1} - a e^{-1} \sum_{j''} d_{j''}^{-1} - a) d_j^{-1} \prod_{j' \neq j} (\lambda - b_{j'} d_{j'}^{-1}).$$

References

Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Statist. 1970; 41:164–171.

Caron F, Doucet A. Efficient Bayesian inference for generalized Bradley-Terry models. 2010 arXiv: 1011.1761.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc Ser B. 1977; 39:1–38. With discussion.

Gromley I, Murphy T. A mixture of experts model for rank data with applications in election studies. Annals of Applied Statistics. 2008; 2:1452–1477.

Haseman JK, Soares ER. 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. 1976; 41:277–288.

Heiser, WJ. Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. Oxford University Press; Oxford: 1995. p. 157-189.

Hines RJO, Lawless JF. Modelling overdispersion in toxicological mortality data grouped over time. Biometrics. 1993; 49:107–121. [PubMed: 8513096]

Hunter DR. MM algorithms for generalized Bradley-Terry models. Ann Statist. 2004; 32:384–406.

Hunter DR, Lange K. A tutorial on MM algorithms. Amer Statist. 2004; 58:30–37.

Ionita-Laza I, Laird NM. On the optimal design of genetic variant discovery studies. Statistical Applications in Genetics and Molecular Biology. 2010; 9:33.

Lange, K. Statistics for Biology and Health. second edition. Springer-Verlag; New York: 2002. Mathematical and Statistical Methods for Genetic Analysis.

Lange, K. Statistics and Computing. second edition. Springer; New York: 2010. Numerical Analysis for Statisticians.

Lange K, Hunter DR, Yang I. Optimization transfer using surrogate objective functions. J Comput Graph Statist. 2000; 9:1–59. With discussion, and a rejoinder by Hunter and Lange.

de Leeuw, J. Information Systems and Data Analysis. Springer; Berlin: 1994. Block-relaxation algorithms in statistics; p. 308-325.

MacKay DJC, Bauman Peto LC. A hierarchical dirichlet language model. Natural Language Engineering. 1994; 1:1–19.

Madsen, RE.; Kauchak, D.; Elkan, C. Modeling word burstiness using the dirichlet distribution. ICML '05 Proceedings of the 22nd international conference on Machine learning; New York, NY, USA: ACM; 2005. p. 545-552.

McLachlan, GJ.; Krishnan, T. Wiley Series in Probability and Statistics. second edition. Wiley-Interscience [John Wiley & Sons]; Hoboken, NJ: 2008. The EM Algorithm and Extensions.

Meng XL. Discussion on "optimization transfer using surrogate objective functions". Journal of Computational and Graphical Statistics. 2000; 9:35–43.

Meng XL, Rubin DB. Using EM to obtain asymptotic variancecovariance matrices: the SEM algorithm. Journal of the American Statistical Association. 1991; 86:899–909.

Paul SR, Balasooriya U, Banerjee T. Fisher information matrix of the Dirichlet-multinomial distribution. Biom J. 2005; 47:230–236. [PubMed: 16389920]

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Computer Applications in the Biosciences: CABIOS. 1996; 12:327–345. [PubMed: 8902360]

Tvedebrink T. Overdispersion in allelic counts and θ-correction in forensic genetics. Theoretical Population Biology. 2010; 78:200–210. [PubMed: 20633572]

Wu TT, Lange K. The MM alternative to EM. Statistical Science. 2010; 25:492–505.

Zhou H, Alexander D, Lange K. A quasi-newton acceleration for high-dimensional optimization algorithms. Statistics and Computing. 2011; 21:261–273. [PubMed: 21359052]

Zhou H, Lange K. MM algorithms for some discrete multivariate distributions. Journal of Computational and Graphical Statistics. 2010; 19:645–665. [PubMed: 20877446]

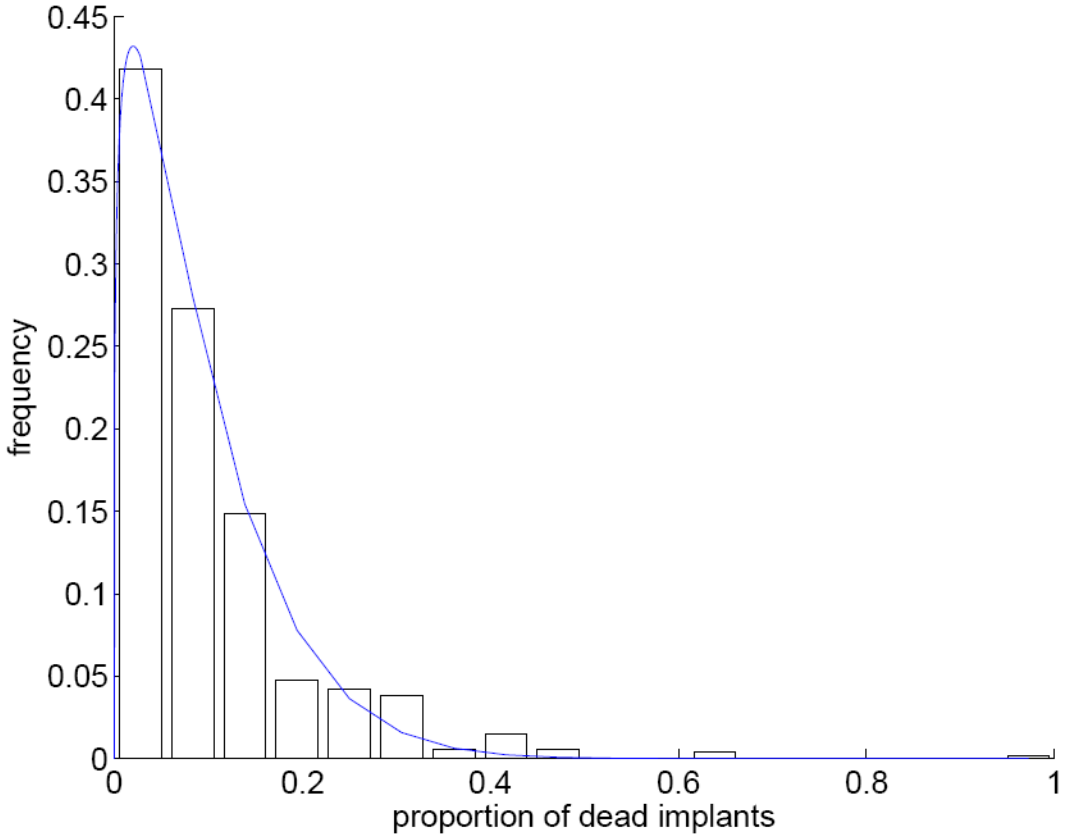

Figure 1.
Histogram of the 524 proportions in the Haseman and Soares data with a Beta(1.23,12.46) density imposed.

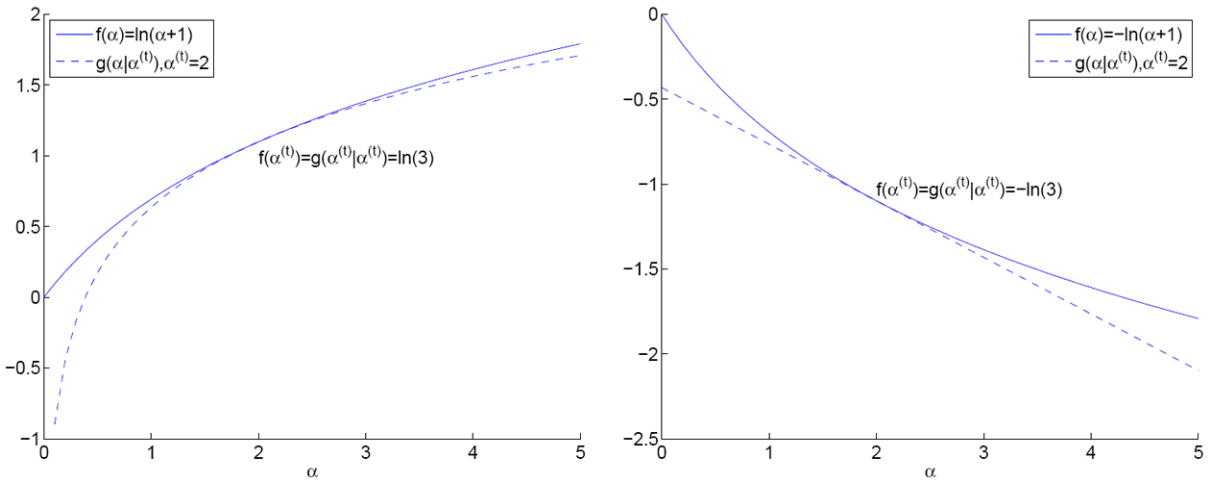

Figure 2.
Graphs of minorization in equalities (7) and (8).

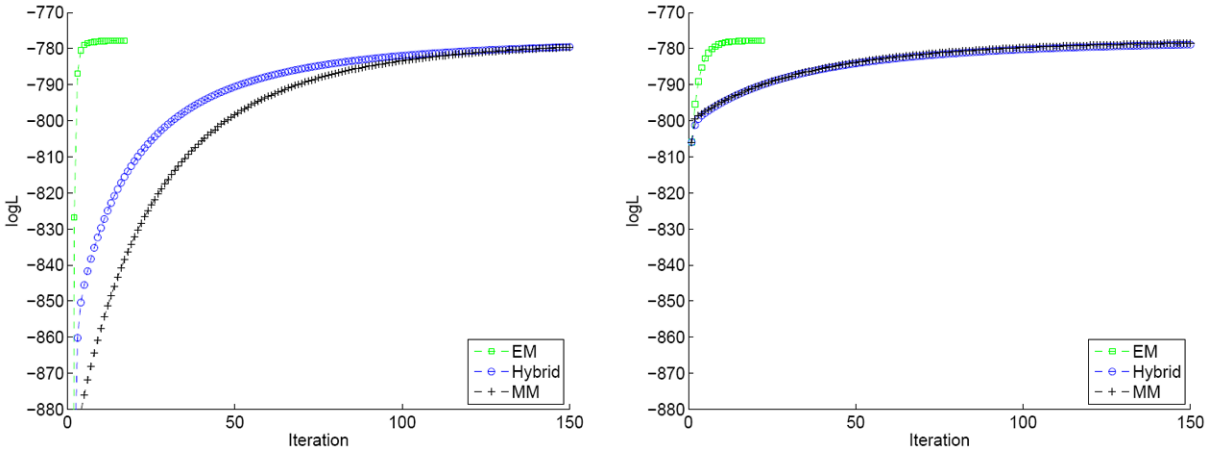

Figure 3.
Algorithmic iterates for the HS76-1 data set. Left: Start from a blind guess $\boldsymbol{\alpha}^{(0)} = (1, 1)$. Right: Start from the method of moment estimate $\boldsymbol{\alpha}^{(0)} = (0.4711, 4.8072)$.

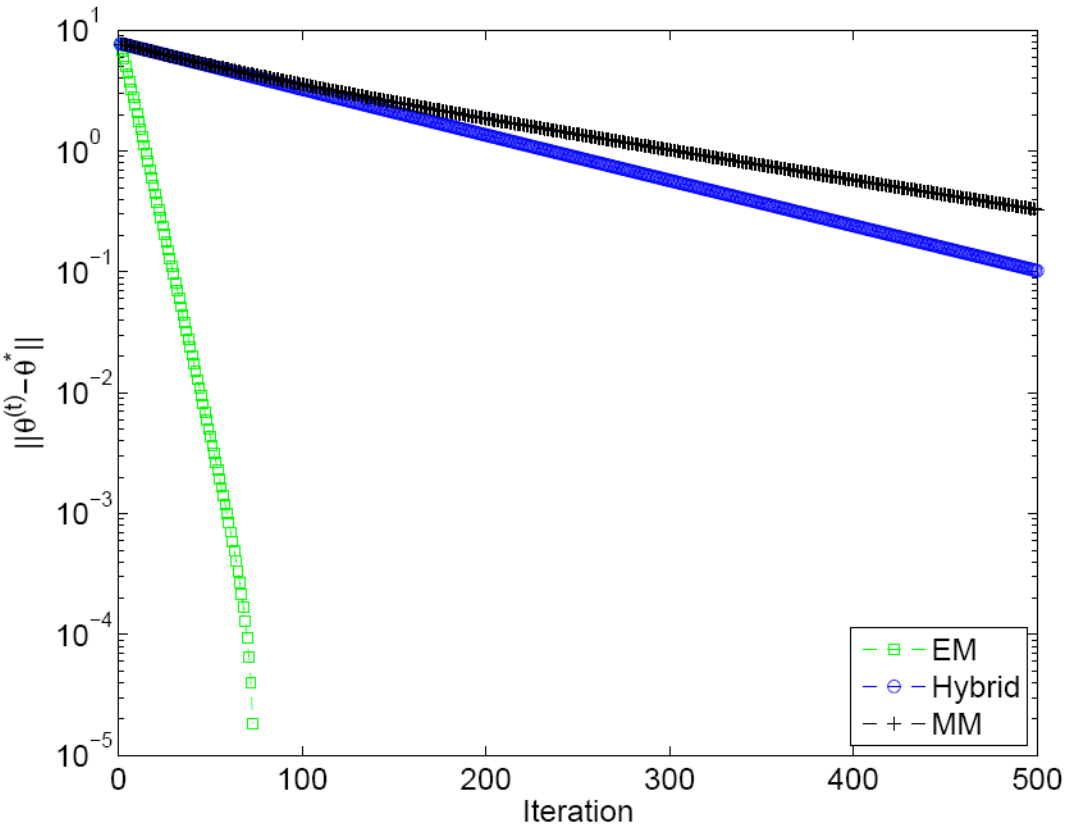

Figure 4.
Distance of algorithmic iterates to the final solution $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{\infty}\|_2$ for the HS76-1 data set.

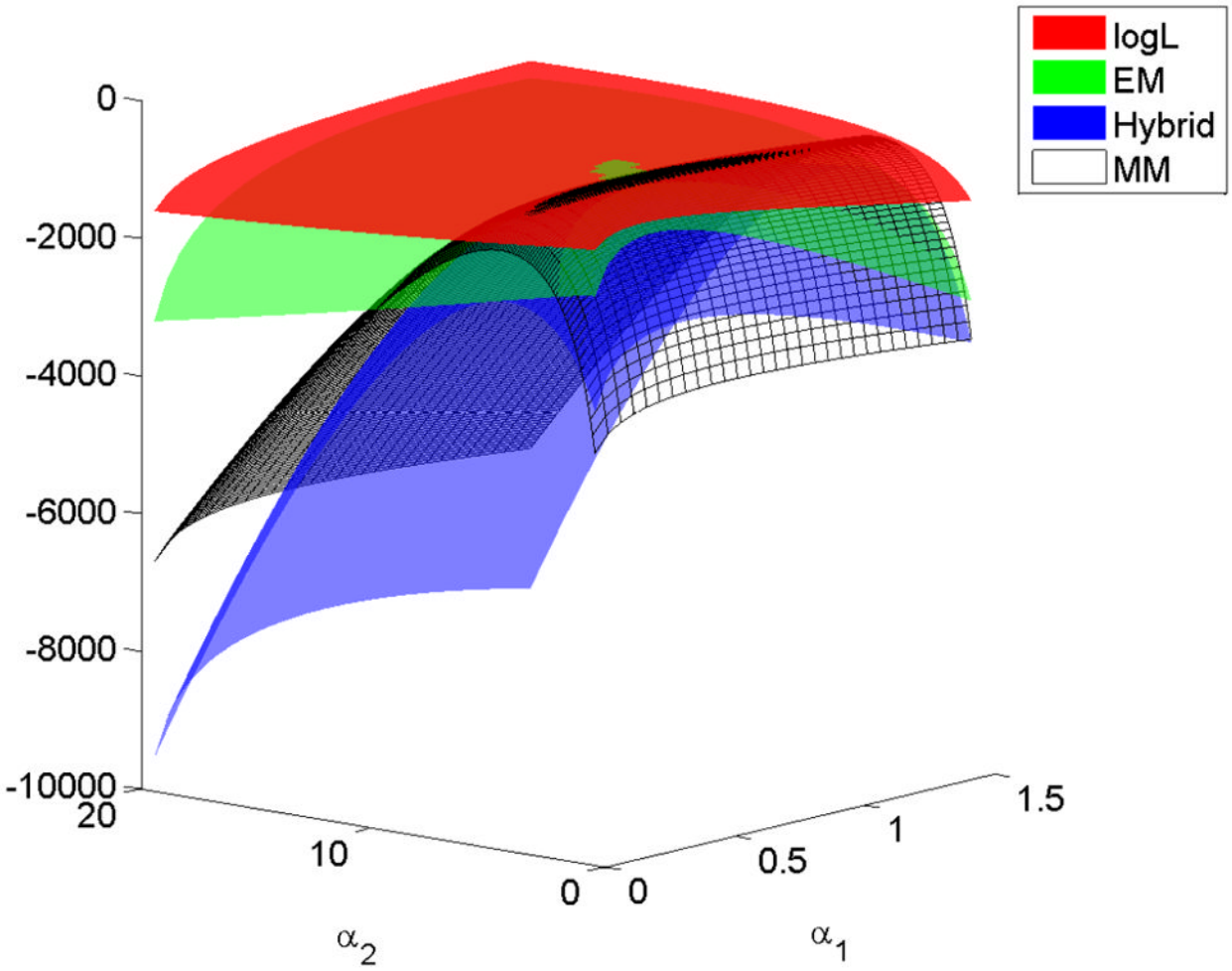

Figure 5.
Log-likelihood surface and the minorizing functions of EM, MM, and the EM-MM hybrid algorithms at point (0.5,5) for the HS76-1 data set.

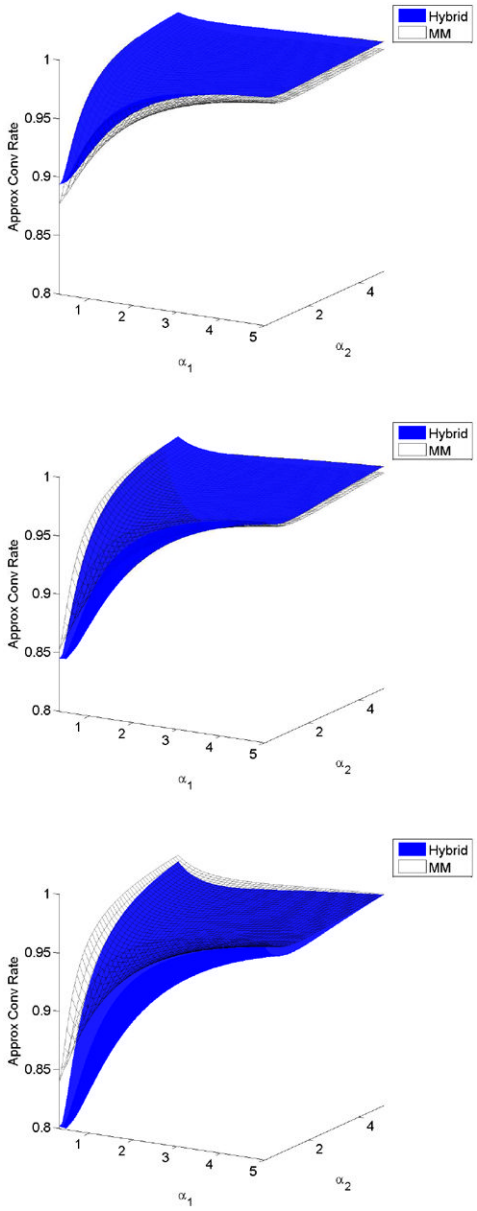

Figure 6.

Approximate convergence rates of the MM and hybrid algorithms for fitting the beta-binomial distribution ($d = 2$) when all data points have the same batch size m. Top: $m = 5$; Middle: $m = 10$; Bottom: $m = 20$.

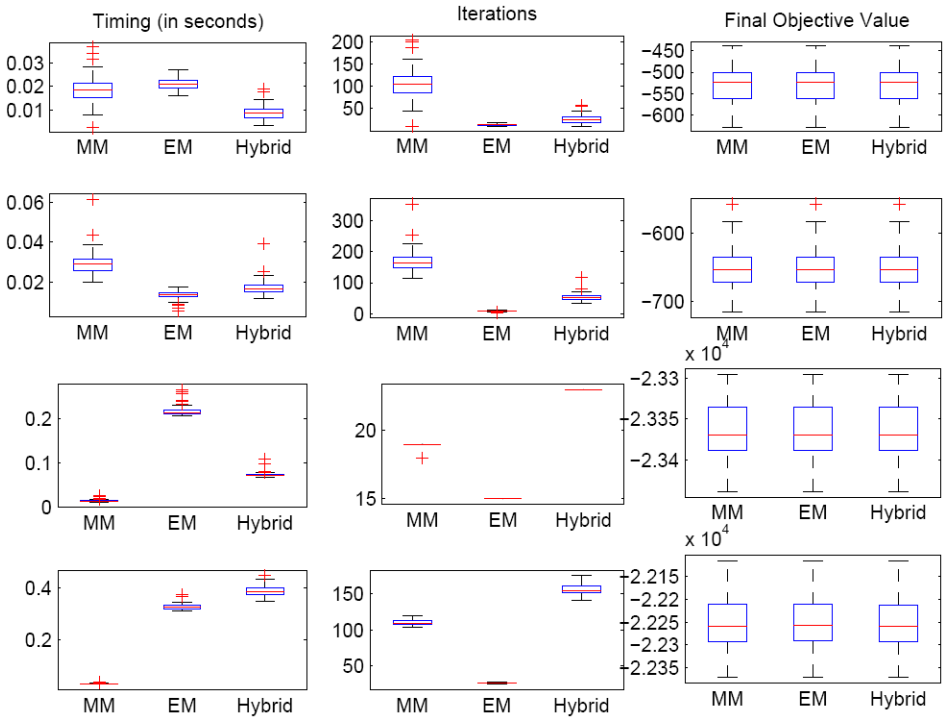

Figure 7.

Comparison of algorithmic timing, convergence rates, and final objective values under different parameter values. Row 1: $\boldsymbol{\alpha} = (0.1, 1)$; Row 2: $\boldsymbol{\alpha} = (0.2, 2)$; Row 3: $\alpha_1 = \cdots = \alpha_{50} = 0.5$; Row 4: $\alpha_1 = \cdots = \alpha_{50} = 5$. The sample size is $n = 200$. The batch size is $m = 20$. There are 100 replicates in each scenario. Convergence criterion is 10^{-6}.