The Multiset EM Algorithm

Weihong Huang and Yuguo Chen*

*Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright
Street, Champaign, IL 61820 USA*

Abstract

The expectation-maximization (EM) algorithm is widely used in computing the maximum likelihood estimates when the observations can be viewed as incomplete data. However, the convergence rate of the EM algorithm can be slow especially when a large portion of the data is missing. We propose the multiset EM algorithm that can help the convergence of the EM algorithm. The key idea is to augment the system with a multiset of the missing component, and construct an appropriate joint distribution of the augmented complete data. We demonstrate that the multiset EM algorithm can outperform the EM algorithm, especially when EM has difficulties in convergence and the E-step involves Monte Carlo approximation.

Keywords: EM algorithm; Data augmentation; Multiset; Incomplete data.

*Telephone: +1 217 244 3932.

Fax: +1 217 244 7190.

Email address: yuguo@illinois.edu.

1 Introduction

The EM algorithm (Dempster et al., 1977), along with many of its variations, such as the MCEM algorithm (Wei and Tanner, 1990), the ECM algorithm (Meng and Rubin, 1993), and the PX-EM algorithm (Liu et al., 1998), are useful tools for finding the maximum likelihood estimates (MLEs) in incomplete data problems. The EM algorithm has some nice properties, including its easy implementation and numerical stability. However, sometimes the EM algorithm is slow to converge. Dempster et al. (1977) showed that the rate of convergence of the EM algorithm is linear and depends on the fraction of information loss due to incompleteness of the data.

There is a close connection between the EM algorithm and the data augmentation strategy proposed by Tanner and Wong (1987). Data augmentation shares the same incomplete-data scheme with the EM algorithm, and it can be viewed as a stochastic generalization of the EM algorithm (Liu and Wu, 1999). A more general data augmentation scheme, called the multiset sampler (MSS), was proposed in Leman et al. (2009). The MSS augments the system with a multiset of the nuisance parameters, and the standard data augmentation is the MSS with multiset size one. The MSS has been shown to improve the convergence of the data augmentation scheme in many examples (Leman et al., 2009).

Because of the close connection between EM and data augmentation, we expect that the multiset idea should be able to be applied to the missing component in the EM algorithm and improve the convergence of the EM algorithm as well. In this paper, we incorporate the multiset idea into EM and develop the multiset EM algorithm. We augment the data with a multiset of the missing component and construct a proper joint distribution on the augmented complete data. Then the standard EM scheme can be implemented on the augmented system. We will show that the multiset EM can aid the convergence of the standard EM in several examples. In practice, although the multiset EM can converge in fewer number of iterations, it may take longer computation time than the standard EM due to the additional computation cost in the E-step. However, when the standard EM has difficulties in convergence and the E-step itself involves Monte Carlo approximation, the multiset EM can outperform the EM algorithm.

The paper is organized as follows. Section 2 describes the multiset EM algorithm. Section 3 discusses about some computation issues about the multiset EM algorithm and gives several numerical examples. Section 4 concludes with a discussion.

2 The Multiset EM

Suppose Y_{obs} is the set of observed data with log-likelihood $\log p(Y_{\text{obs}}|\theta)$ depending on parameter θ . The objective is to find the MLE θ_{MLE} , which maximizes $\log p(Y_{\text{obs}}|\theta)$. When direct maximization of $\log p(Y_{\text{obs}}|\theta)$ over θ is difficult, the EM algorithm augments Y_{obs} to a set of complete data Y_{com} , which has log-likelihood $\log p(Y_{\text{com}}|\theta)$. We write $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{mis} denotes the missing or unobserved data. Starting with an initial guess θ_0 of θ , the EM algorithm gives an iterative way to maximize $\log p(Y_{\text{obs}}|\theta)$ using an expectation (E) step and a maximization (M) step at each iteration. At iteration t , suppose the current value of θ is θ_t , the EM algorithm updates θ_t as follows:

- E-step: Calculate the expected log-likelihood of complete data with respect to the conditional distribution $p(Y_{\text{mis}}|Y_{\text{obs}}, \theta_t)$, i.e., $Q(\theta|\theta_t) = E_{Y_{\text{mis}}|Y_{\text{obs}}, \theta_t}[\log p(Y_{\text{obs}}, Y_{\text{mis}}|\theta)]$.
- M-step: Find θ_{t+1} by maximizing $Q(\theta|\theta_t)$ over θ .

Let $t = t+1$ and repeat the above updating scheme until the convergence criterion is satisfied.

Before describing the multiset EM algorithm, we first review the definition of multiset, following Leman et al. (2009). Multiset is a generalized notion of set, which allows repeats. Suppose $\mathcal{Y}$ is a set, then a multiset of size k on $\mathcal{Y}$ contains k elements of $\mathcal{Y}$. These elements can be identical and the order is irrelevant. For example, $\{1, 2, 1, 3, 1\}$ is a multiset of size 5 on the integer set $\mathbb{N}$. It is the same as $\{1, 1, 1, 2, 3\}$ but different from $\{1, 1, 2, 2, 3\}$. Let Y be a discrete random variable with support $|\mathcal{Y}| < \infty$, and let p be the probability mass function of Y . Then we can define the distribution of the random multiset S with size k on $\mathcal{Y}$ as

$$q(S) = q(y_1, y_2, \dots, y_k) = C^* \sum_{y \in S} p(y),$$

where the normalizing constant

$$C^* = \frac{|\mathcal{Y}|}{k \binom{|\mathcal{Y}|+k-1}{k}}$$

because there are $\binom{|\mathcal{Y}|+k-1}{k}$ such multisets.

Now we introduce the multiset EM algorithm. Let Ω be the support of Y_{mis} , that is, Ω contains all possible values of Y_{mis} . When Y_{mis} is discrete and $|\Omega| < \infty$, let S_k be a random multiset of size k on Ω , and define the joint probability distribution of the augmented complete data $Y_{\text{com},k} = (Y_{\text{obs}}, S_k)$ given θ as

$$p(Y_{\text{com},k}|\theta) = p(Y_{\text{obs}}, S_k|\theta) = C \sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{obs}}, Y_{\text{mis}}|\theta), \quad (1)$$

where C is the normalizing constant. Therefore, we also have

$$p(S_k|Y_{\text{obs}}, \theta) = \frac{p(Y_{\text{obs}}, S_k|\theta)}{p(Y_{\text{obs}}|\theta)} \propto \frac{\sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{obs}}, Y_{\text{mis}}|\theta)}{p(Y_{\text{obs}}|\theta)} = \sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{mis}}|Y_{\text{obs}}, \theta).$$

Under this setting, the multiset EM algorithm can be applied to maximize $\log p(Y_{\text{obs}}|\theta)$. Start with an initial guess θ_0 of θ . At iteration t , suppose the current value of θ is θ_t , the multiset EM algorithm updates θ_t as follows:

- E-step: Calculate the expected log-likelihood of the augmented complete data with respect to the conditional distribution $p(S_k|Y_{\text{obs}}, \theta_t)$, i.e.,

$$Q(\theta|\theta_t) = E_{S_k|Y_{\text{obs}}, \theta_t} [\log p(Y_{\text{com},k}|\theta)] = Q^*(\theta|\theta_t) + \log C,$$

where C is a constant and

$$Q^*(\theta|\theta_t) = E_{S_k|Y_{\text{obs}}, \theta_t} \log \left[\sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \right]. \quad (2)$$

- M-step: Find θ_{t+1} by maximizing $Q^*(\theta|\theta_t)$ over θ .

Let $t = t+1$ and repeat the above updating scheme until the convergence criterion is satisfied.

When the multiset size $k = 1$, the multiset EM algorithm becomes the standard EM algorithm.

3 Numerical Examples

3.1 Computational Issues

When the distribution of the complete data comes from the exponential family, the EM algorithm can compute the E-step analytically, but the multiset EM cannot. The E-step

of the multiset-EM needs to calculate the expected value of the augmented complete data log-likelihood with respect to the conditional distribution $p(S_k|Y_{\text{obs}}, \theta_t)$. If this expectation is hard to compute and the space of the multiset S_k is large, we may use Monte Carlo methods (Wei and Tanner, 1990), such as naive Monte Carlo, important sampling and Markov chain Monte Carlo, to approximate the E-step. For the M-step, if the maximizer of formula (2) has no explicit solution, we use numerical optimization methods.

Compared with the standard EM algorithm, the multiset EM has a higher computation cost due to two reasons. One is the increasing amount of likelihood evaluation when calculating the augmented complete data likelihood in the E-step. The other is the Monte Carlo approximation in the E-step when the space S_k is large. The cost associated with additional likelihood evaluation is often ignorable compared with the cost of the Monte Carlo approximation. Therefore when the E-step in the EM algorithm also needs Monte Carlo approximation, the multiset EM may be beneficial in terms of convergence time.

3.2 Examples

All examples were coded in R and run on a PC with 2.40 GHz Intel Core i7 processor.

3.2.1 A Simple Example

Consider the complete-data model, $Y_1, Y_2 \stackrel{iid}{\sim} \text{Binomial}(n, \theta)$, where $n = 2$ and $\theta \in \{0, 0.01, 0.02, \dots, 1\}$ is the parameter we want to estimate. In this model, $Y_1 = 1$ is observed and Y_2 is missing. It is obvious that the MLE of θ is 0.5. To implement the multiset EM algorithm, we built a multiset S_k with size k on the support of $Y_{\text{mis}} = Y_2$. Then the conditional probability of $S_k = (z_1, \dots, z_k)$ given the observed data Y_{obs} and current estimate $\theta^{(t)}$ is

$$p(S_k|Y_{\text{obs}}, \theta^{(t)}) \propto \sum_{z \in S_k} p(z|Y_{\text{obs}}, \theta^{(t)}),$$

where $p(z|Y_{\text{obs}}, \theta) = p(z|\theta)$ is $\text{Binomial}(2, \theta)$. Therefore the E-step of the multiset EM at iteration t can be computed by

$$Q^*(\theta|\theta^{(t)}) = \sum_{S_k \in \Omega_k} p(S_k|Y_{\text{obs}}, \theta^{(t)}) \log \left[\sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \right],$$

where Ω_k is the set containing all possible S_k 's. For the M-step, since the support of θ is discrete, the grid search method can be applied to find the maximum.

Starting from the initial value $p = 0.1$, the first eight steps of the EM algorithm and the multiset EM with three different choices of multiset size k are displayed in Figure 1. The results show that the multiset EM takes fewer steps to converge compared to EM, and the number of iterations needed to achieve convergence seems to decrease as the multiset size k increases. Figure 2(a) shows the total time until convergence for EM and the multiset EM with different multiset size. This figure indicates that in terms of computation time, the multiset EM takes longer to converge than EM, and the amount of time needed seems to increase with k . This is due to the increasing amount of likelihood calculation in the E-step. Figure 2(b) shows the computation time for each iteration of the multiset EM with different multiset size k . The plot suggests that the computation time for each iteration increases roughly linearly as the multiset size k increases. Note that $k = 1$ corresponds to the standard EM algorithm.

Figure 1: Convergence history of θ for EM and the multiset EM (MEM) with different multiset size k for the simple example.

Figure 2: (a) Total time until convergence for EM and the multiset EM (MEM) with different multiset size k for the simple example. (b) Computation time for each iteration of the multiset EM with different multiset size k for the simple example.

3.2.2 Genetic Linkage Model

The genetic linkage model, which has been examined by Dempster et al. (1977) and Wei and Tanner (1990), is multinomial with four categories with the following cell probabilities

$$p = \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4} \right).$$

Assume that 197 animals are multinomially distributed according to the above model, and the observed data $\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$. We regard $\mathbf{y}$ as incomplete data from a five categories multinomial model with cell probabilities $(1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$. By splitting the first category into two categories, we augment the data as $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, where $x_1 + x_2 = 125$, $x_3 = y_2$, $x_4 = y_3$ and $x_5 = y_4$.

To implement the multiset EM algorithm, let the missing component $Y_{\text{mis}} = x_2$ with support $\Omega = \{0, 1, 2, \dots, 125\}$, and we can build a multiset S_k with size $k = 2$ on the support Ω . The total number of all possible S_k is $|\Omega| \times (|\Omega| + 1)/2 = 126 \times 127/2 = 8001$.

The conditional probability of $S_k = (z_1, z_2)$ given the observed data Y_{obs} and current estimate $\theta^{(t)}$ is

$$p(S_k|Y_{\text{obs}}, \theta^{(t)}) \propto p(z_1|Y_{\text{obs}}, \theta^{(t)}) + p(z_2|Y_{\text{obs}}, \theta^{(t)}),$$

where $p(z|Y_{\text{obs}}, \theta)$ is Binomial($n = 125, p = \frac{\theta/4}{1/2+\theta/4}$). Therefore the E-step of the multiset EM at iteration t can be computed by

$$Q^*(\theta|\theta^{(t)}) = \sum_{S_k \in \Omega_k} p(S_k|Y_{\text{obs}}, \theta^{(t)}) \log \left[\sum_{Y_{\text{mis}} \in S_k} p(Y_{\text{obs}}, Y_{\text{mis}}|\theta) \right],$$

where Ω_k is the set containing all possible S_k 's, and $p(Y_{\text{obs}}, Y_{\text{mis}}|\theta)$ is Multinomial($m = 197, p = (1/2, \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4)$). For M-step, there is no explicit maximizer of $Q^*(\theta|\theta^{(t)})$, so we use numerical solutions at M-step.

To compare the multiset EM with the standard EM algorithm, we applied both algorithms to this genetic linkage data starting from the initial value $\theta^{(0)} = 0.3$. The M-step of multiset EM was computed by the built-in `optimize` function in R. To have a fair comparison, we also used the same numerical algorithm for the M-step of EM. The results are given in Table 1. The table provides the value of $\theta^{(t)}$, the absolute error of the estimate at the current iteration $|\theta^{(t)} - \theta^*|$, where θ^* is the true MLE, and the ratio of successive deviations calculated as $|(\theta^{(t)} - \theta^*)/(\theta^{(t-1)} - \theta^*)|$. A smaller ratio indicates a faster converge rate. In this example, the ratio for the multiset EM is always smaller than the standard EM, which indicates that the multiset EM converges to the true MLE $\theta^* = 0.6268215$ faster than the standard EM in terms of the number of iterations. The EM algorithm takes 0.03 seconds per iteration, while the multiset EM takes 2.6 seconds per iteration due to the large amount of likelihood evaluation in the E-step caused by the large size of Ω_k .

3.2.3 Gaussian Mixture Model

Consider a two-component Gaussian mixture model:

$$Y_1 \sim N(\mu_1, \sigma_1^2),$$

$$Y_2 \sim N(\mu_2, \sigma_2^2),$$

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,$$

Table 1: Comparison between EM and the multiset EM for the linkage data

Iteration	EM			Multiset EM		
	$\theta^{(t)}$	$ \theta^{(t)} - \theta^* $	$\left \frac{\theta^{(t)} - \theta^*}{\theta^{(t-1)} - \theta^*} \right $	$\theta^{(t)}$	$ \theta^{(t)} - \theta^* $	$\left \frac{\theta^{(t)} - \theta^*}{\theta^{(t-1)} - \theta^*} \right $
0	0.300000	0.326822	-	0.300000	0.326822	-
1	0.569668	0.057154	0.174877	0.582039	0.044783	0.137024
2	0.618915	0.007907	0.138341	0.621321	0.005501	0.122829
3	0.625746	0.001076	0.136030	0.626181	0.000640	0.116441
4	0.626698	0.000123	0.114441	0.626763	0.000058	0.090847

where $\Delta \in \{0, 1\}$ and $P(\Delta = 1) = \pi$. We simulated $n = 20$ i.i.d. data points (listed in Table 2) from the mixture normal with two components $N(-5, 2.5^2)$, $N(5, 2.5^2)$ and $\pi = 0.6$. We would like to estimate $\theta = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$.

Table 2: Twenty simulated data points from mixture normal distribution

6.998	7.163	5.877	-7.577	-4.299	10.176	1.517	5.645	7.565	-8.328
2.866	-7.549	3.480	-3.098	5.228	11.371	-1.961	5.517	-4.569	7.761

To estimate θ by the maximum likelihood method, we consider the unobserved latent variable (or missing component) $\Delta_i \in \{0, 1\}$, which takes value 0 if observation Y_i comes from the first component and 1 if Y_i comes from the second component. For each Δ_i , we build a multiset S_{ki} with size $k = 2$. Then there are three possible cases for each S_{ki} , that is $S_{ki} = (z_{i1}, z_{i2})' \in \{(0, 0)', (0, 1)', (1, 1)'\}$. Given the observed data Y_i and current estimate $\theta^{(t)}$, the conditional probability of S_{ki} follows the three-category Multinomial(p), where

$$p \propto \left(2(1 - \pi^{(t)})\phi(Y_i; \mu_1^{(t)}, \sigma_1^{(t)}), (1 - \pi^{(t)})\phi(Y_i; \mu_1^{(t)}, \sigma_1^{(t)}) + \pi^{(t)}\phi(Y_i; \mu_2^{(t)}, \sigma_2^{(t)}), 2\pi^{(t)}\phi(Y_i; \mu_2^{(t)}, \sigma_2^{(t)}) \right),$$

and $\phi(Y_i; \mu, \sigma)$ is the density function of $N(\mu, \sigma^2)$ evaluated at Y_i .

Let $S_k = (S_{k1}, S_{k2}, \dots, S_{kn})$ and Ω_k be the set which contains all possible S_k 's. Then the

E-step of the multiset EM at iteration t should be computed as

$$Q^*(\theta|\theta^{(t)}) = \sum_{S_k \in \Omega_k} \left(\prod_{i=1}^n p(S_{ki}|Y_i, \theta^{(t)}) \right) \log \left[\prod_{i=1}^n \sum_{z \in S_{ki}} p(Y_i, z|\theta) \right],$$

where

$$p(Y_i, z|\theta) = [(1 - \pi)\phi(Y_i; \mu_1, \sigma_1)]^{(1-z)} \cdot [\pi\phi(Y_i; \mu_2, \sigma_2)]^z.$$

However, the set Ω_k includes 3^n components, which makes the exact computation of $Q^*(\theta|\theta^{(t)})$ difficult. Instead, we use naive Monte Carlo method to approximate it.

We applied both the EM algorithm and the multiset EM to this problem starting from the same initial value $\theta^{(0)} = (\pi^{(0)} = 0.3, \mu_1^{(0)} = 0, \mu_2^{(0)} = 0.5, \sigma_1^{(0)} = \sigma_2^{(0)} = 2)$. The M-step of the multiset EM was computed by sequentially optimizing each parameter by using the built-in `optimize` function in R. At the beginning of the iterations, 2000 Monte Carlo samples were used to approximate $Q^*(\theta|\theta^{(t)})$ in the E-step, and in the later iterations, 5000 samples were used to obtain more accurate approximation. The EM and multiset EM were stopped when $\|\theta^{(t)} - \theta^{(t-1)}\|_\infty = \max_i |\theta_i^{(t)} - \theta_i^{(t-1)}| < 0.1$. The convergence history of π and μ_1 is shown in Figure 3. We can see that under the same stopping criterion, the multiset EM converged faster than EM in terms of the number of iterations. The EM algorithm needs about 20 iterations to converge while the multiset EM needs only 13 iterations. In terms of computation time, the multiset EM takes more time to converge than the EM algorithm, because it uses Monte Carlo approximation in the E-step (which is the most time-consuming part) while the EM algorithm does not. When the E-step of the EM algorithm can be computed analytically, the multiset EM often takes longer to implement due to the computational difficulty with its E-step. When the EM algorithm also requires Monte Carlo in the E-step, the multiset EM can converge faster in terms of computation time. An example is given in the next section.

3.2.4 The SEIR Model

The SEIR model is a type of epidemic model which has been used to study the mechanisms of the transmission of infectious diseases through individuals. In the SEIR model, the population can be classified into four stages: susceptible, exposed, infectious, and recovered.

Figure 3: Convergence history of π (left) and μ_1 (right) for EM and the multiset EM (MEM) algorithms for the Gaussian mixture example.

We use (S_t, E_t, I_t, R_t) to denote the number of people in each stage at time t . In this section, we study a simple version of the SEIR model described in Leman et al. (2009).

The model assumes that at time t , people can only stay at their current stage, move to the next stage, or die. We use notation such as N_t^{SE} to denote the number of people transferred from stage S at time t to stage E at time $t + 1$. We assume that the population size N stays the same, so that we have the same birth rate and death rate (denoted by μ for both rates). Let p denote the percentage of people who are born with natural immunity or vaccinated at an early age. Then the model can be described as follows:

$$\begin{aligned} S_{t+1} &= S_t + (1 - p)\mu N - N_t^{SE} - \mu S_t, \\ E_{t+1} &= E_t + N_t^{SE} - N_t^{EI} - \mu E_t, \\ I_{t+1} &= I_t + N_t^{EI} - N_t^{IR} - \mu I_t, \\ R_{t+1} &= R_t + N_t^{IR} - \mu R_t + p\mu N, \end{aligned}$$

where $N_t^{SE} \sim \text{Binomial}(S_t, 1 - \exp(-\beta I_t/N))$, $N_t^{EI} \sim \text{Binomial}(E_t, 1 - \exp(-\sigma))$, and $N_t^{IR} \sim \text{Binomial}(I_t, 1 - \exp(-\gamma))$.

We simulated data from the model above with parameters $\mu = 0.01$, $p = 0.1$, $\sigma = 1/6$,

$\gamma = 1/8$, and $\beta = 0.7$. The initial states were set as $S_1 = 200$, $E_1 = 200$, $I_1 = 1,000$, $R_1 = 8,600$ with total population $N = 10,000$. We simulated 2,000 time steps and used the last 50 time points of N_t^{EI} , $\{N_t^{EI}; t = 1951, \dots, 2000\}$, as the observed data. The parameters μ , p , σ , γ were assume known. We want to find the MLE of β .

We applied both EM and the multiset EM to this example. The missing components include S_{1951} , E_{1951} , I_{1951} , and $\{N_t^{SE}, N_t^{IR}\}_{t=1951}^{2000}$. When applying the multiset EM, we augmented the system by $k = 2$ replicates of the missing components. For both EM and the multiset EM, the E-step requires Monte Carlo approximation. For the EM algorithm, we compute the E-step by the Metropolis-Hastings (M-H) algorithm with the following steps:

1. Change S_{1951} , E_{1951} , I_{1951} sequentially by 1 or -1 with probability $1/2$. Accept or reject each new value according to the M-H rule.
2. Randomly choose a time t from 1951 to 2000, and change N_t^{SE} , N_t^{IR} sequentially by 1 or -1 with probability $1/2$. Accept or reject the new value according to the M-H rule.

For the multiset EM algorithm, we first randomly selected one multiset component, and then implemented the same M-H steps above.

Let $\beta_0 = 0.9$. For both algorithms, we ran $n = 20,000$ M-H iterations at the beginning stage and $n = 200,000$ iterations at the end. To save the computation time and reduce the dependence between M-H samples, after the first 10,000 burn-in iterations, we only used 2,000 equally spaced samples from the remaining iterations to estimate the E-step. The M-steps for both EM and the multiset EM were obtained using the built-in `optimize` function in R. The convergence history of β for both methods is shown in Figure 4. Both EM and the multiset EM converge to about 0.744, and the multiset EM seems to converge faster compared to EM. To determine the convergence of these two algorithms, we set the convergence criterion as $|\beta_t - 0.744| < 0.01$ and $|\beta_{t+1} - \beta_t| < 10^{-4}$. According to this criterion, the multiset EM converged in 77 steps while the standard EM converged in 132 steps. Moreover, the total computation time until convergence for the multiset EM is only 60.5% of that for the standard EM. Therefore in this example, the multiset EM converges faster than EM in terms of not only the number of iterations, but also the total computation time.

Figure 4: Convergence history of β for EM and the multiset EM (MEM) algorithms for the SEIR model.

4 Discussion

We proposed the multiset EM algorithm to improve the performance of the EM algorithm. The joint distribution constructed on the augmented multiset flattens the margin of the missing component. So it removes some of the information contained in the missing part and enhances the fraction of the observed information with respect to the complete information. Therefore the multiset EM improves the convergence rate. More studies about the convergence property of the multiset EM algorithm can be found in Huang (2017), in which the multiset EM is shown to have a larger global speed of convergence than the EM algorithm. In the examples we have tested, the multiset EM always achieves convergence in fewer number of iterations than EM. On the other hand, the multiset EM usually takes longer to run due to the increased amount of likelihood evaluations and the Monte Carlo approximation in the E-step. When the E-step of the EM algorithm involves Monte Carlo approximation and the algorithm has difficulties in convergence, the multiset EM can outperform EM in

terms of both the number of iterations and the computation time, as demonstrated in the SEIR model in Section 3.2.4.

In most of the examples, we only considered the multiset with size $k = 2$. For the MSS, some examples in Leman et al. (2009) suggest that $k = 2$ is large enough. However, the examples in the earlier version of the MSS, proposed by Leman et al. (2007), indicate that the MSS could converge to the target distribution faster as the multiset size increases. This seems to agree with the results of the simple example in Section 4.1 where the multiset EM seems to converge faster as k increases. On the other hand, as k increases, the number of possible multisets increases as well, which will lead to an increase in computational cost. In practice, we suggest choosing a small multiset size.

Our current algorithm requires that the missing component is discrete and has a finite support, which is similar to the requirement of the MSS (Leman et al., 2009). This algorithm also works if the missing component is continuous and lives on a bounded interval, which is another situation considered in Leman et al. (2009). Kim and MacEachern (2015) proposed a way to generalize the MSS to continuous infinite space, and this idea can be applied to the multiset EM algorithm to make it more general.

We briefly describe the framework of the generalized multiset EM here. Let Ω be the support of Y_{mis} and S_k be a random k -tuple on Ω (a traditional vector or array with size k whose elements are from Ω). Let $S_k = (Y_{\text{mis}}^{(1)}, \dots, Y_{\text{mis}}^{(k)})$. Define the joint probability distribution of the augmented complete data $Y_{\text{com},k} = (Y_{\text{obs}}, S_k)$ given θ as

$$p(Y_{\text{com},k}|\theta) = p(Y_{\text{obs}}, S_k|\theta) = \sum_{i=1}^k \left(\alpha_i p(Y_{\text{obs}}, Y_{\text{mis}}^{(i)}|\theta) \prod_{j \neq i} f_j(Y_{\text{mis}}^{(j)}) \right), \quad (3)$$

where α_i 's are multiset weights satisfying $\sum_{i=1}^k \alpha_i = 1$, and $f_i(\cdot)$'s are instrumental densities which are usually probability densities with support equal to Ω . Then we have

$$\begin{aligned} p(S_k|Y_{\text{obs}}, \theta) &= \frac{p(Y_{\text{obs}}, S_k|\theta)}{p(Y_{\text{obs}}|\theta)} = \frac{\sum_{i=1}^k \left(\alpha_i p(Y_{\text{obs}}, Y_{\text{mis}}^{(i)}|\theta) \prod_{j \neq i} f_j(Y_{\text{mis}}^{(j)}) \right)}{p(Y_{\text{obs}}|\theta)} \\ &= \sum_{i=1}^k \left(\alpha_i p(Y_{\text{mis}}^{(i)}|Y_{\text{obs}}, \theta) \prod_{j \neq i} f_j(Y_{\text{mis}}^{(j)}) \right). \end{aligned} \quad (4)$$

Under this setting, the generalized multiset EM algorithm can be applied to maximize

$\log p(Y_{\text{obs}}|\theta)$. Start with an initial guess θ_0 of θ . At iteration t , suppose the current value of θ is θ_t , the generalized multiset EM algorithm updates θ_t as follows:

- E-step: Calculate the expected log-likelihood of the augmented complete data with respect to the conditional distribution $p(S_k|Y_{\text{obs}}, \theta_t)$, i.e.,

$$Q(\theta|\theta_t) = E_{S_k|Y_{\text{obs}}, \theta_t} [\log p(Y_{\text{com},k}|\theta)],$$

where $p(Y_{\text{com},k}|\theta_t)$ and $p(S_k|Y_{\text{obs}}, \theta_t)$ are given in (3) and (4) respectively.

- M-step: Find θ_{t+1} by maximizing $Q(\theta|\theta_t)$ over θ .

Let $t = t+1$ and repeat the above updating scheme until the convergence criterion is satisfied.

There are some tuning parameters in the generalized multiset EM algorithm, such as the multiset weights α_i and the instrumental densities f_i . Kim and MacEachern (2015) discussed the choice of these tuning parameters for the generalized MSS, and those suggestions are useful for the generalized multiset EM algorithm as well. The E-step in the generalized multiset EM can be hard to compute since $p(S_k|Y_{\text{obs}}, \theta_t)$ is more difficult to evaluate or sample and the complete likelihood takes longer time to compute. Further research on the generalized multiset EM will be useful.

Acknowledgements

This work was supported in part by National Science Foundation grant DMS-1406455.

References

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Huang, W. (2017), “Statistical Algorithms Using Multisets and Statistical Inference on Heterogeneous Networks,” PhD Dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

- Kim, H. J. and MacEachern, S. N. (2015), “The generalized multiset sampler,” *Journal of Computational and Graphical Statistics*, 24, 1134–1154.
- Leman, S. C., Chen, Y., and Lavine, M. (2009), “The multiset sampler,” *Journal of the American Statistical Association*, 104, 1029–1041.
- Leman, S. C., Uyenoyama, M. K., Lavine, M., and Chen, Y. (2007), “The evolutionary forest algorithm,” *Bioinformatics*, 23, 1962–1968.
- Liu, C., Rubin, D. B., and Wu, Y. N. (1998), “Parameter expansion to accelerate EM: the PX-EM algorithm,” *Biometrika*, 85, 755–770.
- Liu, J. S. and Wu, Y. N. (1999), “Parameter expansion for data augmentation,” *Journal of the American Statistical Association*, 94, 1264–1274.
- Meng, X.-L. and Rubin, D. B. (1993), “Maximum likelihood estimation via the ECM algorithm: A general framework,” *Biometrika*, 80, 267–278.
- Tanner, M. A. and Wong, W. H. (1987), “The calculation of posterior distributions by data augmentation,” *Journal of the American statistical Association*, 82, 528–540.
- Wei, G. C. and Tanner, M. A. (1990), “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms,” *Journal of the American Statistical Association*, 85, 699–704.