On Single Versus Multiple Imputation for a Class of Stochastic Algorithms Estimating Maximum Likelihood

Edward H. Ip

Marshall School of Business, University of Southern California, Los Angeles, CA90089

Summary

We discuss a special class of stochastic versions of the EM algorithms. The advantage of the single imputation procedure in non-exponential family applications is highlighted. We prove that ergodic properties of the stochastic algorithms are dependent not on the multiplicity of the imputation scheme but rather on the stability of the deterministic component of an underlying stochastic difference equation.

Keywords: EM algorithm, non-exponential family, Markov chain, stochastic difference equation

1 Introduction

Dempster, Laird and Rubin (1977) introduce the Expectation-Maximization (EM) algorithm as a general-purpose iterative method for exact maximum likelihood (ML) estimation in incomplete-data problems. A number of important extensions have since been developed. One of the most notable is the class of stochastic versions, including the Stochastic EM (SEM, Celeux and Diebolt 1985; see also Diebolt and Ip 1996), the Stochastic Approximation EM (SAEM, Delyon, Lavielle, and Moulines 1999), the Monte Carlo EM
(MCEM, Wei and Tanner 1990), and the simulated EM (Ruud 1991) algorithms. Recently, Wang and Robins (1998) discuss the asymptotic behavior of the SEM, MCEM, and simulated EM with multiple imputations. They make an implicit ergodic assumption that the iterative algorithms converge to stationary distributions.

The stochastic versions of EM often aim to circumvent the problem of high-dimensional numerical integration that is required by the E-step. In contrast to the EM algorithm, only approximate ML solutions are available from this class of algorithms.

This note aims to clarify certain distinctive advantages of the SEM algorithm, which uses a single imputed value in its E-step, as compared to other multiple imputation schemes. An issue that arises from SEM is the possible large fluctuations of the Markov chain it generates. We prove an ergodicity result showing that the convergence properties of the Markov chain is not dependent upon whether single or multiple imputation is used. The result provides a basis for applying SEM to situations where a full exploration of the posterior distribution of model parameter is either computationally too expensive or not necessary.

2 Stochastic Versions of the EM Algorithm

Let \( x \sim p(\theta) \), where \( p(\theta) \) is a density function characterized by the parameter \( \theta \in \mathbb{R}^k \). Suppose \( x \) can be partitioned into two components \((y, z)\), where \( y \) denotes the observed data and \( z \) denotes the unobserved or latent data. Define the function \( Q \) as

\[
Q(\theta, \theta') = \int_Z [\log p(\theta | y, z)] p(z | \theta', y) dz,
\]

where \( \log p(\theta | y, z) \) is the complete-data loglikelihood function, \( p(z | \theta', y) \) is the conditional predictive distribution of \( z \) at \( \theta' \), \( z \) current approximation of the parameter \( \theta \), and \( Z \) is the sample space of \( z \). That \( Q(\theta, \theta') \) can be written as

\[
Q(\theta, \theta') = L(\theta) + H(\theta, \theta') - \kappa(\theta'),
\]

(1)

where \( L(\theta) = \log p(\theta | y) \) is the observed loglikelihood function, \( H(\theta, \theta') = \int_Z \log p(z | \theta, y)p(z | \theta', y) dz \), and \( \kappa(\theta') = \int_Z \log p(z | y)p(z | \theta', y) dz \).

In ML estimation, \( \theta \) is estimated by the value of \( \theta \) that maximizes \( L(\theta) \). This maximization often involves a cumbersome high dimensional integral. The idea behind EM is to maximize \( Q(\theta, \theta') \) instead of \( L(\theta) \). The E-step is defined by computing \( Q(\theta, \theta_t) \), where \( \theta_t \) is the current approximation to the maximizer of the observed loglikelihood function at the \( t \)-th iteration, and the M-step is defined by updating the maximizer of \( Q(\theta, \theta_t) \) with respect to \( \theta \), to \( \theta_{t+1} \).
The E-step may or may not have a closed form. In the event that it does not have a closed form solution, in order to avoid performing high-dimensional integration, the method of Monte Carlo can be used to obtain a sample approximation to $Q(\theta, \theta_t)$ in the E-step, namely,

$$Q_t^* (\theta, \theta_t) = \frac{1}{m} \sum_{j=1}^{m} \log(p(\theta|z^*_j, y)),$$

where $\theta_t$ is the estimate for $\theta$ at the t-th iteration, and $z^*_j$ is a sample drawn from the conditional predictive density $p(z|\theta_t, y)$.

The M-step proceeds to maximize the right hand side of (2). The conditional predictive density can be updated using the new maximizer of $Q_t^*$. The algorithm is then iterated. Typically, the stochastic algorithms terminate when successive approximations to the maximizer of the loglikelihood function are close. The value of $\theta$ at which the algorithm terminates, or some ergodic average, is computed and used as an approximation to the ML estimate.

In both the MCEM and SAEM, the algorithm is usually implemented by drawing $m > 1$ simulated samples from the conditional predictive distribution of the latent data. The SEM algorithm, on the other hand, always uses $m = 1$. Although the computational overhead is primarily the same whether $m = 1$ or $m > 1$, using a large number of simulated samples to approximate the E-step changes the form of the original complete data loglikelihood function $\log p(\theta|x, y)$ when the complete data do not belong to an exponential family. An example of non-exponential family applications to data arising from a censored Weibull distribution is described in Diebolt and Ip (1990).

There are two problems when the form of the likelihood equation changes for multiple imputation. First, for the M-step closed form solution may not exist for $m > 1$, even when such solution exists for $m = 1$. Second, when $m > 1$, the M-step maximizes a mixture-type function in (2). It is well-known that maximization of mixture distributions is computationally burdensome and unstable because multiple modes may be present in the function. Therefore, for non-exponential families, the SEM algorithm would have a computational advantage. By using $m = 1$, the SEM algorithm also lends itself to better interpretability because the stochastic E-step can be viewed as an imputation procedure that "fills in" the latent component by a simulated draw of $z$ at the current approximation of the predictive distribution.

It may be argued that an alternative method is to use the mean of $m$ simulated samples as an imputed value for $z$ in the stochastic E-step at each iteration. However, it can be seen from (2) that such a procedure results in biased estimates.

One possible problem with the SEM algorithm is that the noise level associated with a single imputation is higher than with multiple imputations. This may lead to erratic behavior of the Markov chain $\{\theta_t\}$ generated by the SEM algorithm. Hence, a critical question is how does the choice of $m$
affect the ergodic properties of \( \{\theta_t\} \). In the next section we prove an ergodic result concerning \( \{\theta_t\} \). Our result indicates that the ergodic properties of the stochastic EM algorithm are not dependent on the multiplicity of the imputation scheme. This implies that using \( m = 1 \) will not affect the convergence property of the stochastic algorithm. In practice, convergence of the Markov chain still needs to be verified. Methods for Markov Chain Monte Carlo are generally applicable. For a recent discussion, see Chauveau and Diebolt (1999) and the reference therein.

3 Ergodic result

The SEM algorithm generates a sequence of estimates \( \{\theta_t\} \) in which \( \theta_{t+1} \) is a function of \( \theta_t \), and some random error that arises from the simulation of the complete data. The function is difficult to deal with in full generality. Nevertheless, in large samples, we expect that the SEM sequence with \( m = 1 \) to approximately satisfy a simpler stochastic difference equation.

Let \( M : \mathbb{R}^k \rightarrow \mathbb{R}^k \) denote the EM mapping function. That is, \( M(\hat{\theta}_t) = \hat{\theta}_{t+1} \) where \( \{\hat{\theta}_t\} \) is the EM sequence of estimates. Assume that \( M \) is smooth, i.e., continuously differentiable up to an appropriate order. We expect in SEM \( \theta_{t+1} \) to be close to the value of the EM mapping function on \( \theta_t \) plus a random variate:

\[
\theta_{t+1} \approx M(\theta_t) + \frac{S(\theta_t)}{\sqrt{n}} \varepsilon_{t+1},
\]

where \( n \) is the sample size of the observed data \( y \). The positive semi-definite \( k \times k \) matrix \( S(\theta) \) is some smooth function in \( \theta \) and the sequence \( \{\varepsilon_t\} \) is independently distributed \( N(0, I) \), where \( I \) is the identity matrix. Equation (3) is a conditional relation where \( \theta_t \) and \( y \) are both treated as fixed. The noise \( \varepsilon_{t+1} \) that arises in the system is due to the randomness in \( z^* \).

Formally, we show the result in (3) is true under certain regularity conditions. Let \( Q_t \) and \( Q_t^n \) respectively denote \( Q(\theta, \theta_t) \) and its Monte Carlo approximation \( Q_t^n(\theta, \theta_t) \). Suppose \( \theta_{t+1} = \text{arg max} Q_t^n = \varphi(Q_t^n) \) and \( \varphi \) is a smooth function with respect to \( \theta \), and assume that \( z^* \) contains independent and identically distributed samples \( (z_1^*, \ldots, z_r^*) \). Further, assume that the ratio of sample size of the latent data \( z \), \( r \), to the sample size of \( y \), \( n \), is \( r \).

That is, \( r/n = r > 0 \).

**Theorem 1.** Under regularity conditions,

\[
n^{1/2}(\theta_{t+1} - M(\theta_t)) \rightarrow N(0, S^T(\theta_t)S(\theta_t)).
\]

**Proof.** We observe that for \( m = 1, Q_t^n = L(\theta) + \log p(z^*(\theta_t)|\theta, y) - \kappa(\theta_t) \). As \( r \rightarrow \infty, (1/r) \log p(z^*(\theta_t)|\theta, y) \rightarrow (1/r)H(\theta, \theta_t) \) in probability. To see this, note that

\[
H(\theta, \theta_t) = \int \log p(z|\theta, y)p(z|\theta_t, y)dz
\]
\[
\begin{align*}
&= \int \sum_{i=1}^{r} \log p(z_i|\theta, y) p(z_i|\theta_i, y) dz_i \\
&= r \int \log p(z|\theta, y) p(z|\theta_i, y) dz.
\end{align*}
\]

Under the condition that \( \int \log^2 p(z|\theta, y) p(z|\theta_i, y) dz < \infty \), 
\( (1/r) \sum_i \log p(z_i|\theta, y) \rightarrow \int \log p(z|\theta, y) p(z|\theta_i, y) \) in probability, where \( z_i \sim p(z|\theta_i, y) \).

Hence, \( Q_t^*/r \rightarrow Q_t/r \) in probability. Because \( \varphi(Q_t^*) = \varphi(Q_t^*/r) \) for any \( r > 0 \), we have \( \varphi(Q_t^*/r) \rightarrow \varphi(Q_t/r) \). From large sample theory, \( r^{1/2}(Q_t^*/r - Q_t/r) \rightarrow N(0, A) \) under regularity conditions (Lehmann 1983, p. 415, 428), where \( A = A(\theta_t) \) is the asymptotic variance of \( Q_t^*/r \) given \( \theta_t \). By the delta method we have
\[
n^{1/2}(\varphi(Q_t^*) - \varphi(Q_t)) = n^{1/2}(\theta_t + M(\theta) - M(\theta_t)) \rightarrow N(0, r^{-1} \nabla^\top \varphi(Q_t/r) A \nabla \varphi(Q_t/r)) .
\]

When \( m > 1 \), the variance of \( A \) is approximately reduced to \( (1/m) A \).

Hence, for large sample, we use the following the stochastic difference equation for the \( m \)-imputation procedure:
\[
\theta_{t+1} = M(\theta_t) + \frac{S(\theta_t)}{\sqrt{nm}} \varepsilon_{t+1} . \tag{4}
\]

A stochastic system such as that defined by (4) may exhibit irregular behavior for a general function \( M(\theta) \). See, for example, Kifer (1988). In the following, we assume that \( M(\theta) \) is continuous. The following theorem provides conditions under which the Markov chain \( \{\theta_t\} \) converges.

**Theorem 2.** If the Markov chain \( \{\theta_t\} \) generated by the process (4), whereby the \( \{\varepsilon_t\} \) are independent and identically distributed with respect to a continuous positive density but not necessarily normal, is aperiodic, irreducible, and satisfies the following conditions:

(C1) There exists a compact set \( C \) such that \( \|M(\theta)\| \leq \rho \|\theta\| \), where 0 < \( \rho < 1 \) for \( \theta \not\in C \),

(C2) For some constant \( K > 0 \), \( \|S(\theta)\|^2 \leq K \|\theta\|^2 \) for \( \theta \not\in C \),

where \( \|\cdot\| \) denotes Euclidean norm, then \( \{\theta_t\} \) is geometric ergodic and has finite second moment.

An important implication of the theorem is that the ergodic properties of \( \{\theta_t\} \) depend largely on the behavior of \( M(\theta) \). Under conditions (C1) and (C2), the choice of \( m \) does not play a role in determining ergodicity.

The proof is based on an application of the results in Tweedie (1975, 1983); See also Meyn and Tweedie (1993). In the following proof, we use the results of Tweedie given in the form of two lemmas in Chan (1990). They are stated below with slight modifications.

**Lemma 1.** Let \( \{\theta_t\} \) be an aperiodic and irreducible Markov chain. Assume that the transition kernel has a positive continuous density. If there
exists a compact set \( C \), a non-negative measurable function \( g \), constants \( 0 < \rho < 1, \gamma > 0 \), and \( B > 0 \) such that
\[
E(g(\theta_{t+1}) \mid \theta_t = \theta) < \rho g(\theta) - \gamma, \quad x \notin C,
\]
(5)
and
\[
E(g(\theta_{t+1}) \mid \theta_t = \theta) < B, \quad x \in C,
\]
(6)
then \( \{\theta_t\} \) is geometric ergodic.

The function \( g(\theta) \) is regarded as a generalized energy function. The inequality (5) asserts that, if the Markov chain starts outside the compact set \( C \), it would, on average, dissipate energy in the next step. \( C \) can be regarded as the center of the state space.

**Lemma 2.** Let \( \{\theta_t\} \) be an ergodic Markov chain on \( \mathbb{R}^k \) with limiting continuous probability distribution \( \Psi, \Psi(\mathbb{R}^k) = 1 \), and \( f \) is a non-negative measurable function. If for some compact set \( C \) (w.r.t. \( \Psi \)),

\[ f \]

is bounded away from zero and infinity on \( C \),

\[
sup_{C} E[f(\theta_{t+1}) \mid \theta_t = \theta] < \infty, \quad \theta \in C,
\]

there exists \( 0 < \rho < 1 \) such that \( E[f(\theta_{t+1}) \mid \theta_t = \theta] \leq \rho f(\theta), \quad \theta \notin C \),

then
\[
\int_{\mathbb{R}^k} f(\theta)\Psi(d\theta) < \infty.
\]

**Proof of theorem 2.** From (3), the transition kernel has a positive and continuous density. This gives
\[
(\theta_{t+1})^T(\theta_{t+1}) = [M(\theta_t)]^T[M(\theta_t)] + \frac{1}{nm}\varepsilon_{t+1}^T S^T(\theta_t)S(\theta_t)\varepsilon_{t+1} + \frac{2}{\sqrt{nm}}M^T(\theta_t)S(\theta_t)\varepsilon_{t+1}.
\]
Conditional on \( \theta = \theta_t \) and taking expectation with respect to the distribution of \( \varepsilon \) gives
\[
E(||\theta_{t+1}||^2 \mid \theta_t) = ||M(\theta_t)||^2 + \frac{1}{nm}E[\varepsilon_{t+1}^T S^T(\theta_t)S(\theta_t)\varepsilon_{t+1} \mid \theta_t].
\]
(7)
Thus, under assumption (C1) and (C2), for \( \theta_t \notin C \),
\[
E(||\theta_{t+1}||^2 \mid \theta_t) \leq \rho ||\theta_t||^2 + \frac{K}{nm} ||\theta_t||^2
\]
(8)
for some \( 0 < \rho' < 1 \) when \( n \) is large enough. Using Lemma 1 with \( g(x) = ||x||^2 \), condition (5) is met. From (7), observe that inside the compact set \( C \), \( ||M(\theta_t)||^2 \) is bounded uniformly as \( M \) is continuous. The second term on the
right hand side of (7) is \( \text{trace}(S^T(\theta)S(\theta)) / nm \). This term is also uniformly bounded in \( C \) since the trace of the matrix \( S^T(\theta)S(\theta) \) is a continuous function in \( \theta \). Thus \( E[||\theta_{t+1}||^2 | \theta_t] \) is bounded uniformly in the compact set \( C \). As a result, condition (6) in Lemma 1 is met and geometric ergodicity is proved.

To prove existence of the second moment, we apply Lemma 2. First, take \( f(\theta) = ||\theta||^4 + 1 \) in order to have \( f \) bounded away from 0. The second condition in Lemma 2 is met because \( C \) is compact. Finally, the third condition in Lemma 2 is satisfied by virtue of (9). Therefore, we have \( \int (1 + ||\theta||^4) \psi(d\theta) < \infty \), implying that the second moment exists. \( \Box \)

Acknowledgement

This paper is a modified version of a chapter from the author's doctoral dissertation. The author thanks both Jean Diebold and Ingram Olkin for their guidance, and for suggestions to this paper.

Reference


