A STOCHASTIC DIFFERENCE EQUATION APPROACH TO INFERENICE WITH MISSING DATA: SOME NEW RESULTS

by

E.H.S. Ip
J. Diebold

Technical Report No. 303
October 1994

Prepared Under the Auspices of National Science Foundation Grant DMS 93-01366 Ingram Olkin, Project Director

Department of Statistics Stanford University Stanford, California
A STOCHASTIC DIFFERENCE EQUATION APPROACH
TO INFERENCE WITH MISSING DATA:
SOME NEW RESULTS

by

Eddie H.S. Ip
Jean Diebolt

Technical Report No. 303
October 1994

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 93-01366
Ingram Olkin, Project Director

Also supported by Sandia National Laboratory

Department of Statistics
Stanford University
Stanford, California
A Stochastic difference equation approach to inference with missing data: some new results

Eddie H.S. Ip and Jean Diebolt

October 1994

Summary

This paper is concerned with the study of the properties of the Stochastic EM estimator. The Stochastic EM estimator is an estimator derived from an iterative algorithm which handles statistical model with missing data. The first result we obtain is on the decomposition of a general stochastic difference equation underlying the Stochastic EM iterates into an additive form. Such decomposition allows a further analysis of the Markov chain underlying the Stochastic EM algorithm. The second result indicates that the Stochastic EM estimator differs from the maximum likelihood estimate by $O(1/n)$ where $n$ is the sample size. It also shows that iterates derived from the additive stochastic difference equation converges to a normal distribution which is centered at the true value of the parameter as $n \to \infty$. 
A Stochastic difference equation approach to inference with missing data: some new results

Eddie H.S. Ip$^1$ and Jean Diebolt$^2$

1 Introduction

This paper is a continuation to Diebolt and Ip(1994). It concerns with the study of the properties of an estimator, the Stochastic EM estimator. This estimator approximates the maximum likelihood estimate (MLE) in a missing data situation. The motivation for such an estimate is given in Diebolt and Ip(1994).

The procedure that derives the estimate is quite straightforward to state. Denote the observed data by $y$, the missing data by $z$ and the complete data by $x$. So $x = (y, z)$. Suppose the model that we base on is a parametric one with parameter $\theta$. Further suppose that statistical inference such as finding the MLE based on $x$ can be easily performed while inference just based on $y$ is difficult. The procedure begins by giving an initial guess $\theta^{(0)}$ for the parameter $\theta$. Fill in the missing data with a draw $z^{(0)}$ from the conditional distribution $z \mid y, \theta^{(0)}$. Compute the MLE $\theta^{(1)}$ of $\theta$ based on this pseudo-complete sample. Update the parameter with $\theta^{(1)}$ and iterate the above simulation step. As a result this scheme generates dual Markov chains $\{z^{(m)}\}$ and $\{\theta^{(m)}\}$ that would converge to their respective stationary distributions under regular conditions. The proposed estimate is the mean $\bar{\theta}$ of the stationary distribution of $\theta$. Since the procedure is analogous to the EM algorithm (Dempster et al.,1977), we call $\bar{\theta}$ the Stochastic EM estimate and the procedure outlined above the Stochastic EM algorithm. This paper is a preliminary investigation of the properties of $\bar{\theta}$. The MLE is denoted by $\hat{\theta}$ throughout this paper.

Section 2 gives several examples of the stationary distribution of the chain generated from $\{\theta^{(m)}\}$. These are special cases in which the stationary distribution can be derived explicitly. In general, the stationary distribution has

$^1$Stanford University
$^2$Universite Paris-6
no analytic form. Section 3 describes how the Markov chain \( \{ \theta^{(m)} \} \) can be characterized by a stochastic difference equation. The ergodicity of the chain is treated in Section 4. Finally Section 5 and 6 are devoted to the analysis of the Markov chain \( \{ \theta^{(m)} \} \) for the linear and non-linear case respectively.

2 Examples

Example 2.1 Missing at random.

Consider a simple example in which certain values are randomly missing in random variables generated from a normal distribution with unknown mean \( \theta \) and known variance \( \sigma^2 \). Denote the observed data by \( y_1, \ldots, y_r \) and the missing data by \( z_1, \ldots, z_k \). Let \( n = r + k \). Since the data points generated are independent, the conditional density of \( z_i \mid y, \theta^{(m)} \) is \( \mathcal{N}(\theta^{(m)}, \sigma^2) \). Suppose that \( z_1^{(m)}, \ldots, z_k^{(m)} \) are drawn from this distribution at the m-th iterate to complete the sample, so that we have an update of \( \theta \) given by

\[
\theta^{(m+1)} = (r\bar{y} + k\bar{z}^{(m)})/n = a\theta^{(m)} + b\bar{y} + \sqrt{kn}^{-1}\sigma \varepsilon^{(m+1)},
\]

where \( \bar{y} = \frac{1}{r} \sum_{i=1}^{r} y_i \), \( a = k/n \), \( b = 1 - a \), and \( \varepsilon^{(m+1)} \sim \mathcal{N}(0,1) \) independently. Indeed, the Markov process \( \{ \theta^{(m)} \} \) is driven by the transition kernel, \( K(d\theta \mid \theta, y) = P(\theta^{(1)} \in d\theta \mid \theta^{(0)} = \theta, y) \) where \( d\theta \) represents the density of the Gaussian distribution

\[
\mathcal{N}(a\theta + b\bar{y}, \frac{k}{n^2}\sigma^2).
\]

The stationary distribution of the Markov chain of estimates generated by Stochastic EM is \( \mathcal{N}(\bar{y}, k\sigma^2/r(n + k)) \) and \( \tilde{\theta} = \bar{y} = \hat{\theta} \).

Example 2.2 Contaminated signal.

Suppose that a Gaussian signal \( \{ Z_i \} \), \( i = 1, \ldots, n \), is contaminated by a Gaussian white noise and we are only able to observe the contaminated signal \( \{ Y_i \} \). Formally, let \( Z_i \sim \mathcal{N}(\theta, \sigma^2_1) \), and \( Y_i = Z_i + \sigma_2 \varepsilon_i \), where \( \varepsilon_i \sim \mathcal{N}(0,1) \) for all \( i \). In other words, \( Y_i \mid Z_i \sim \mathcal{N}(Z_i, \sigma^2_2) \). The parameter of interest is \( \theta \) and we assume that \( \sigma_1 \) and \( \sigma_2 \) are known. It is immediate that \( \hat{\theta} = \bar{y} \). Treat the signal \( Z_i \) as missing values and apply the Stochastic EM algorithm. It is straightforward to show that the stationary distribution of \( \theta \) is

\[
\mathcal{N}\left(\bar{y}, \frac{\sigma^2_2(\sigma^2_1 + \sigma^2_2)}{n(\sigma^2_1 + 2\sigma^2_2)}\right).
\]
So again \( \hat{\theta} = \hat{\theta} \).

**Example 2.3. Contaminated bivariate signal.**

We extend the above example to the bivariate case. Suppose that for \( i = 1, \ldots, n \), \( Z_i \sim \mathcal{N}(\theta, \Sigma_1) \), and \( Y_i \mid Z_i \sim \mathcal{N}(Z_i, \Sigma_2) \), where \( Z_i = (Z_{i1}, Z_{i2})^T \).

Assume that

\[
\Sigma_1 = \begin{pmatrix}
\sigma^2 & \rho \sigma^2 \\
\rho \sigma^2 & \sigma^2
\end{pmatrix}
\]

and \( \Sigma_2 = \tau^2 I_2 \). Suppose further that all parameters are known except for \( \theta \).

The conditional distribution of \( \theta^{(m+1)} \) given \( Y \) and \( \theta^{(m)} \) is \( \mathcal{N}(A\theta^{(m)} + B\bar{y}, (1/n)A\Sigma_1) \), where \( A = I - \Sigma_1(\Sigma_1 + \Sigma_2)^{-1} \) and \( B = \Sigma_1(\Sigma_1 + \Sigma_2)^{-1} \). In order to find the mean \( \mu \) and variance \( \Gamma \) of the stationary distribution, which is known to be Gaussian, we need to solve a set of equations. For the mean \( \mu \), the fixed point equation is \( \mu = A\mu + B\bar{y} \) and the solution to is \( \mu = \bar{y} \).

For the variance \( \Gamma \), the fixed point equation is

\[
\Gamma = \frac{1}{n} A \Sigma_1 + A \Gamma A^T.
\] (2.1)

Denote \( \tau^2 / \sigma^2 \) by \( r \), the variance of the conditional distribution is

\[
\frac{1}{n} A \Sigma_1 = \frac{1}{n}(\Sigma_1 - \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}\Sigma_1)
= \frac{\sigma^2 r}{n[(1 + r)^2 - \rho^2]} \begin{pmatrix}
1 + r + \rho^2 & 2\rho + r \\
2\rho + r & 1 + r + \rho^2
\end{pmatrix}.
\]

Now write

\[
\Gamma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{pmatrix}
\]

in (2.1) and solve for \( \sigma_{11}, \sigma_{22}, \sigma_{12} \). If \( r \) is small, we can ignore the \( r^2 \) terms and obtain the following approximations to the variance

\[
\sigma_{11} = \sigma_{22} \approx \frac{\sigma^2 r (1 + \rho^2)}{n(1 + 2r - \rho^2)},
\]

and the correlation

\[
\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} \approx \frac{2\rho + r}{1 + r + \rho^2} \leq 1.
\]

Note that the variance increases when \( \rho \) gets close to 1. An increase in variance can be attributed to a loss of information when the components in
the signal are highly correlated. This brings about extremely slow convergence when EM is applied. Such a phenomenon is investigated in Thomas and Ip (1994) in greater detail. They also provide examples with real data from the National Assessment of Educational Progress (NAEP).
3 Simplifying the Stochastic Difference Equation

The Stochastic EM algorithm has a close tie to the EM algorithm. Denote the estimate of $\theta$ at the m-th iterate in EM by $\theta^{(m)}$ and let $T$ be the operator that maps $\theta^{(m)}$ to $\theta^{(m+1)}$. That is, $T(\theta^{(m)}) = \theta^{(m+1)}$. The properties of the stationary distribution derived from the Stochastic EM algorithm, when it exists, depends very much on the EM operator $T(.)$. In general, the Stochastic EM iterates take the form of a stochastic difference equation

$$\theta^{(m+1)} = f(\theta^{(m)}, \varepsilon^{(m+1)}(\theta^{(m)})),$$

which is difficult to solve in full generality. This section explains how (3.1) can be decomposed into a more tractable form

$$\theta^{(m+1)} = T(\theta^{(m)}) + V(\theta^{(m)})\varepsilon^{(m+1)},$$

when the complete data is from an exponential family.

We follow closely the notation in Dempster et al. (1977) and Wu (1983). Let the density function of the complete data $x$ be $f(x | \theta)$ and that of $y$ be

$$g(y | \theta) = \int_y f(x | \theta) \, dx.$$ 

where $\mathcal{Y} = \{x : y(x) = y\}$. Also let $k(x | y, \theta) = f(x | \theta) / g(y | \theta)$ be the conditional density of $x$ given $y$ and $\theta$. Write the loglikelihood equation as

$$\ell_c(\theta; y, z) = \ell_{\text{cond}}(\theta; y, z) + \ell_{\text{obs}}(\theta; y).$$

where $\ell_c$ denotes the complete data loglikelihood, $\ell_{\text{cond}}$ the conditional loglikelihood of the complete data given the observed data and $\ell_{\text{obs}}$ the observed loglikelihood. Suppose now a sample $z_o$ is imputed for the missing data.

There are two MLE's based on two different loglikelihood functions: the complete data loglikelihood and the conditional loglikelihood. Denote the MLE using the complete data loglikelihood by $\hat{\theta}^c$ and the MLE based on the conditional loglikelihood $\log k(z | y)$ by $\hat{\theta}^k$. In other words,

$$\hat{\theta}^c = \arg \max_{\theta} \ell_c(\theta; y, z_o),$$
\[ \hat{\theta}^k = \arg \max_{\theta} \ell_{\text{cond}}(\theta; z_o \mid y). \]

For instance, in Example 2.1, \( \hat{\theta}^c = a\bar{z} + b\bar{y} \) and \( \hat{\theta}^k = \bar{z} \). To emphasize the dependence of both MLE’s on \( z_o \), we use \( \hat{\theta}^c(z_o) \) and \( \hat{\theta}^k(z_o) \) when necessary.

It is natural at this point to introduce continuity and differentiability assumptions before stating the first lemma. Henceforth in this paper we assume the existence and continuity of a sufficient number of derivatives whenever Taylor expansion is invoked. We also assume that expectation and differentiation can be interchanged.

Suppose the complete data is generated from an exponential family of the form

\[ f(x \mid \theta) = \exp[<\theta, t(x)> - A(\theta)]h(x), \tag{3.4} \]

where \( \theta \in \Theta \) is a parameter. Both \( x \) and \( \theta \) are multidimensional in general, \( \theta \) and \( t(x) \) are vectors of common dimension \( k \), and \( <\cdot, \cdot> \) denotes inner product. For the pseudo-complete sample \((y, z_o)\), make the following assumption on the sufficient statistics \( t(x) = t(y, z_o) \),

\[ t(y, z_o) = E[t(y, z) \mid \theta^*, y] \text{ for some } \theta^* \in \Theta. \tag{3.5} \]

**Lemma 1** Under (3.5), \( T(\hat{\theta}^k(z_o)) = \hat{\theta}^c(z_o) \).

**Proof.** Under (3.5), \( t(y, z_o) = E[t(y, z) \mid \theta^*, y] \) for some \( \theta^* \). Regard \( t(y, z_o) \) as the value of the sufficient statistics obtained during a particular iterate of EM, starting from the guess \( \theta^* \).

It is clear then that the M-step gives \( \hat{\theta}^c(z_o) \) as the update. Therefore it suffices to prove that \( \theta^* \) is necessarily \( \hat{\theta}^k(z_o) \). Once this is established, combining the E-step and the M-step on \( \hat{\theta}^k(z_o) \) would yield \( \hat{\theta}^c(z_o) \), which is what we want.

To prove \( \theta^* = \arg \max_{\theta} \ell_{\text{cond}}(\theta; z_o \mid y) \), first note that when the full data loglikelihood is from an exponential family with sufficient statistics \( t(y, z) \), the conditional distribution of \( z \) given \( y \) also has sufficient statistics which are linear in the loglikelihood and has the same form as in the unconditional case except that now \( y \) is regarded as fixed. This is due to the fact that the conditional density of \( z \) given \( y \) is proportional to the joint density of \((y, z)\). The difference of the loglikelihood of the conditional and the joint is the logarithm of the normalizing term which is not dependent upon \( z \).
Using this fact and the well known fact that the expectation of the score function is zero (Lehmann, 1988; Dempster et al., 1977, Lemma 2), we have, for any \( \theta \),

\[
E \left( \frac{\partial}{\partial \theta} \ell_{\text{cond}}(\theta; t(y, z)) \mid \theta, y \right) = 0.
\]

Hence, for that particular \( \theta^* \) that satisfies (3.5)

\[
E \left( \frac{\partial}{\partial \theta} \ell_{\text{cond}}(\theta^*; t(y, z)) \mid \theta^*, y \right) = 0.
\]

Exchanging expectation and differentiation and using linearity of the sufficient statistics in the conditional distribution, we have

\[
\frac{\partial}{\partial \theta} \ell_{\text{cond}}(\theta^*; E(t(y, z) \mid \theta^*, y)) = 0,
\]

\[
\frac{\partial}{\partial \theta} \ell_{\text{cond}}(\theta^*; t(y, z_o)) = 0,
\]

which implies that \( \theta^* \) is in fact the MLE based on the conditional distribution since in a regular exponential family, a solution to the likelihood equation exists and is the unique maximizer. ◊

The next theorem uses Lemma 1 to obtain the decomposition (3.2).

**Theorem 1** Suppose the complete data \( x = (y, z) \) consists of \( n \) random sample drawn from (3.4). Let the number of sample points in \( z \) be \( r \), the EM operator be \( T \) and \( \theta^{(m)} \) be the value of the parameter at the \( m \)-th iterate of Stochastic EM. If the following conditions holds,

- (A.1) (3.5) is satisfied for a sufficiently large \( m \).
- (A.2) As \( n \to \infty, r/n \to \tau \) where \( 0 < \tau < 1 \),
- (A.3) \( T(w) = T(\theta) + DT(\theta)(w - \theta) + o(||w - \theta||) \), as \( w \to \theta \),

where \( ||.|| \) denotes Euclidean norm and \( DT \) denotes the Jacobian of \( T \), then as \( n \to \infty \), the distribution of \( \theta^{(m+1)} - T(\theta^{(m)}) \) tends to the normal distribution

\[
N(0, DT(\theta^{(m)})J_{\text{cond}}^{-1}(\theta^{(m)})DT(\theta^{(m)})),
\]

where \( J_{\text{cond}} \) is the Fisher information matrix of \( \ell_{\text{cond}} \).
Proof. At the m-th iterate of Stochastic EM, the current guess of the parameter is \( \theta^{(m)} \). For imputation, a sample \( z_n \) is drawn from the conditional distribution of \( z \) given \( y \) and \( \theta^{(m)} \). Under classical likelihood theory, \( \hat{\theta}^k - \theta^{(m)} \) is asymptotically normal with mean 0 and variance \( J^{-1}_{\text{cond}}(\theta^{(m)}) \) as \( r \to \infty \) (Rao, 1973; Lehmann, 1988). By Lemma (1), \( \theta^{(m+1)} = T(\theta^k) \). Under assumption A.2 and using the multivariate delta method of distribution (Bishop, Fienberg, and Holland, 1975, p. 493), the asymptotic distribution of \( (\theta^{(m+1)} - T(\theta^{(m)})) \) is normal with mean 0 and variance 
\[
\text{var}(\theta^{(m+1)}) = (DT)J^{-1}_{\text{cond}}(DT)^T
\]
as \( n \to \infty \) (\( r \to \infty \) if and only if \( n \to \infty \)). The matrices \( DT \) and \( J_{\text{cond}} \) are evaluated at \( \theta^{(m)} \). ♦

Corollary 1 For any parameter \( \eta = \phi(\theta) \), where \( \phi(.) \) is continuous and one-to-one, under the same conditions (A.1), (A.2) and (A.3) in Theorem 1, as \( n \to \infty \), the distribution of \( \eta^{(m+1)} - T_\eta(\eta^{(m)}) \) tends to the normal distribution with mean 0 and variance \( DT(\eta^{(m)})J_{\text{cond}}(\eta^{(m)})DT(\eta^{(m)}) \).

Proof. Note that \( T_\eta = (\phi \circ T) \) is the new EM operator. The corollary is a direct result of applying the delta method for distributions. ♦

The variance of the stationary distribution can actually be related to the information matrices by using the next lemma. Taking the second derivative of (3.3), multiply by \(-1\) and then take expectation with respect to the conditional distribution \( k(x \mid y, \theta) \) yields 
\[
E \left[ -\frac{\partial^2}{\partial \theta^2} \ell_c(\theta; y, z) \mid y, \theta \right] = E \left[ -\frac{\partial^2}{\partial \theta^2} \ell_{\text{cond}}(\theta; y, z) \mid y, \theta \right] + \frac{\partial^2}{\partial \theta^2} \ell_{\text{obs}}(\theta; y).
\]
(3.7)

We can interpret equation (3.7) as "complete information" equals "conditional information" plus "observed information". To simplify notation, write (3.7) as 
\[
J_c = J_{\text{cond}} + J_{\text{obs}}.
\]
(3.8)

Lemma 2 If the regularity conditions

- \( T(\theta^{(m)}) \) converges to \( \hat{\theta} \) in the closure of \( \Theta \) as \( m \to \infty \),
\begin{itemize}
\item \( \frac{\partial}{\partial \theta} E(\ell_c(\theta; x) \mid \theta^{(m)}, y) = 0 \) at \( \theta^{(m+1)} \),
\item \( -\frac{\partial^2}{\partial \theta \partial \theta} E(\ell_c(\theta; x) \mid \theta^{(m)}, y) \) is positive definite with eigenvalues bounded away from 0,
\end{itemize}

holds, then \( DT(\hat{\theta}) = J_{\text{cond}}(\hat{\theta}) J_c^{-1}(\hat{\theta}) \).

See Dempster et. al (1977, Theorem 4) for a proof. The conditions generally hold for regular exponential families.

**Corollary 2** Under the setting of Theorem 1, if \( \theta \) is a scalar and \( J_c \) and \( J_{\text{cond}} \) are independent of \( \theta \), then \( \theta^{(m+1)} - T(\theta^{(m)}) \) tends to \( \mathcal{N}(0, J_c^{-1} J_{\text{cond}} J_c^{-1}) \).

**Proof.** By Lemma (2)

\[
( DT ) J_{\text{cond}} ( DT )^T = ( J_{\text{cond}} J_c^{-1} ) J_{\text{cond}}^{-1} ( J_{\text{cond}} J_c^{-1} )
= J_c^{-1} J_{\text{cond}} J_c^{-1}.
\]

The result is a consequence of Theorem 1. \( \diamond \)

**Remark:**

1. \( J_c^{-1} \) is \( O(1/n) \), \( J_{\text{cond}} \) is \( O(r) \). Because \( r/n \to \tau > 0 \), \( J_c^{-1} J_{\text{cond}} J_c^{-1} \) is \( O(1/n) \).

2. Lemma (2) gives interpretable quantification of the variance of the stationary distribution. This lemma will be used again in a more general setting in the last section.

We have accomplished, at least in an asymptotic way, a decomposition of of (3.1) into a more tractable form (3.2). Rewrite (3.2) into the following form:

\[
\theta^{(m+1)} = T(\theta^{(m)}) + \frac{S(\theta^{(m)})}{\sqrt{n}} \varepsilon^{(m+1)},
\]

(3.9)

where \( \theta^{(m)} \) is a \( k \times 1 \) vector, \( S(\theta) \) is a symmetric positive definite \( k \times k \) matrix, each entry being a function of \( \theta \), \( \varepsilon^{(m+1)} \) is a \( k \times 1 \) random vector independent and identically distributed from \( \mathcal{N}(0, I) \) for \( m = 1, 2, \ldots \).

To obtain results on the stationary distribution of the Markov chain generated by \( \{\theta^{(m)}\} \), first assume that the functions \( T(.) \) and \( S(.) \) are continuous and has an appropriate number of continuous derivatives. In addition, assume that the Markov chain generated from (3.9) aperiodic and irreducible (Chan,1990,p.456). These properties are required in proving ergodicity.
4 Ergodicity

Tweedie(1975,1983) obtain important ergodic results on Markov chain of the kind we are considering. We state two theorems adapted with slight modifications from Chan(1990).

**Theorem 2** Let \( \{\theta^{(m)}\} \) be aperiodic and \( \lambda \)-irreducible. Assume that the transition kernel has a positive continuous density. If there exists a compact set \( C \), a non-negative measurable function \( g \), and constants \( 0 < \rho < 1 \), \( \gamma > 0 \), and \( B > 0 \) such that

\[
E(g(\theta^{(m+1)}) | \theta^{(m)} = \theta) < \rho g(\theta) - \gamma, \quad x \notin C, \tag{4.1}
\]

and

\[
E(g(\theta^{(m+1)}) | \theta^{(m)} = \theta) < B, \quad x \in C, \tag{4.2}
\]

then \( \{\theta^{(m)}\} \) is geometrically ergodic.

As remarked in Chan(1990), \( g(\theta) \) is regarded as a generalized energy function. The inequality (4.1) 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.

**Theorem 3** Let \( \{\theta^{(m)}\} \) be an ergodic Markov chain on \( \mathbb{R}^k \) with limiting probability \( \Psi, \Psi(\mathbb{R}^k) = 1 \), and \( f \) is a non-negative measurable function. If for some compact set \( A \) (w.r.t. \( \Psi \)),

- \( f \) is bounded away from zero and infinity on \( A \),
- \( \sup E[f(\theta^{(m+1)}) | \theta^{(m)} = \theta] < \infty, \quad \theta \in A, \)
- there exists \( 0 < \rho < 1 \) such that
  \[
  E[f(\theta^{(m+1)}) | \theta^{(m)} = \theta] \leq \rho f(\theta), \quad \theta \notin A; \tag{4.3}
  \]

then

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

**Theorem 4** For the dynamic process (3.9), suppose that the chain is aperiodic, irreducible and satisfies the conditions
• (E.1) There exists a compact set $C$ such that $||T(\theta)|| \leq \rho||\theta||$, $0 < \rho < 1$ for $\theta \notin C$,

• (E.2) For some constant $K > 0$, $||S(\theta)||^2 \leq K||\theta||^2$;

then $\{\theta^{(m)}\}$ is geometrically ergodic and has finite second moments.

Proof. From (3.9), the transition kernel has a positive and continuous density. Multiply both sides of (3.9) by their transposes. This gives

$$(\theta^{(m+1)})^T(\theta^{(m)}) = [T(\theta^{(m)})^T[T(\theta^{(m)})]] + \frac{1}{n} \varepsilon^T S^T(\theta^{(m)}) S(\theta^{(m)}) \varepsilon \varepsilon \\
+ 2 < T(\theta^{(m)}), S(\theta^{(m)}) \varepsilon / \sqrt{n} >, \tag{4.4}$$

where $< , , >$ denotes an inner product. For simplicity, the superscript on $\varepsilon$ has been dropped. Conditioning on $\theta^{(m)}$ and taking expectation with respect to $\varepsilon$ gives

$$||\theta^{(m+1)}||^2 = ||T(\theta^{(m)})||^2 + \frac{1}{n} E[\varepsilon^T S^T(\theta^{(m)}) S(\theta^{(m)}) \varepsilon]. \tag{4.5}$$

Thus, under assumption (E1) and (E2), for $\theta \notin C$,

$$E\{||\theta^{(m+1)}||^2 \mid \theta^{(m)}\} \leq \rho ||\theta^{(m)}||^2 + \frac{K}{n} ||\theta^{(m)}||^2 \tag{4.6}$$

< $\rho' ||\theta^{(m)}||^2$, \tag{4.7}

for some $0 < \rho' < 1$ when $n$ is large enough. Using Theorem 2 with $g(x) = ||x||^2$, condition (4.1) is met. From (4.5), first observe that inside the compact set $C$, $||T(\theta)||^2$, is bounded uniformly as $T$ is continuous. The second term on the right hand side of (4.5) is $\text{trace}(S^T(\theta)S(\theta))/n$. This term is also uniformly bounded in $\theta$ since the trace of the matrix $S^T(\theta)S(\theta)$ is a continuous function in $\theta$. Thus $E\{||\theta^{(m+1)}|| \mid \theta^{(m)} = \theta\}$ is bounded uniformly in the compact set $C$. As a result, condition (4.2) in Theorem 2 is met and geometric ergodicity is proved.

To prove existence of the second moments, we invoke Theorem 3 and check on the three conditions that lead to the existence results. Take $f(\theta) = ||\theta||^4 + 1$ in order to have $f$ bounded away from 0. The second condition is met because $C(= A$ in Theorem 3) is compact. Finally, the third condition in Theorem 3 is satisfied by virtue of (4.7). Therefore we have $\int (1 + ||\theta||^4) \pi(d\theta) < \infty$, implying that the second moment exists. $\Diamond$
5 Solving the Stochastic Difference Equation: Linear Case

The exact stationary distribution for process (3.9), even when it exists, is difficult to obtain in full generality. This section gives explicit expression in the multivariate case for the mean and variance of the stationary distribution when $T$ is linear and the noise is independent of $\theta^{(m)}$. A similar discussion of the linear case for continuous time can be found in Kunita (1991). Our results for the general $T$ in the multivariate setting are provided under stronger assumptions in Section 6.

Throughout the following exposition, $\rho(X)$ denotes the largest eigenvalue of the symmetric matrix $X$ and $||X|| = [\rho(X^TX)]^{1/2}$ denotes the spectral norm of $X$. Whereas $||z||$ denotes the Euclidean norm when $z$ is a vector.

**Theorem 5** Let $A$ be a $k \times k$ matrix with $0 < ||A|| < 1$. If

$$\theta^{(m+1)} = A\theta^{(m)} + b + \varepsilon^{(m+1)},$$

(5.1)

where $b$, $\theta^{(m)} \in \mathbb{R}^k$ and $\varepsilon^{(m+1)}$ is the random perturbation in the $m$-th iterate of Stochastic EM with $\varepsilon^{(m)} \sim \mathcal{N}(0, V)$ independently, then the stationary distribution of $\Psi(\theta)$ is asymptotically normal with mean $\bar{\theta} = (I - A)^{-1}b$ and variance $\Upsilon = \sum_{i=1}^{\infty} A^{i-1}V(A^T)^{i-1}$, where $||V||(1 - ||A||^2) \geq ||\Upsilon|| > ||V||$.

**Proof.** The stationary distribution is given by the solution of the convolution equation of (5.1),

$$\Psi(\theta) = \int K(\theta | \tau, y)d\Psi(\tau),$$

(5.2)

where $K(\theta | \tau, y)$ is the Gaussian kernel with mean $A\tau + b$ and variance $V$. When $K$ is normal, it can be shown that the solution $\Psi$ is also a normal distribution by direct verification. Formally, the stationary distribution $\Psi$ can be determined by Fourier transform methods. Recall that the characteristic function of a normal distribution with mean $\mu$ and covariance matrix $\Upsilon$ is $\exp(it^T\mu)\exp[-(1/2)t^T\Upsilon t]$. Our claim is that $\Psi$ is a normal distribution whose mean $\mu$ and variance $\Upsilon$ can be determined by substituting the characteristic function of $\Psi$ into (5.2): The convolution equation (5.2) can
be solved by equating the respective characteristic functions of distributions in
\[ \exp(it^T \mu - \frac{1}{2} t^T \Sigma t) = \exp(it^T A \mu - \frac{1}{2} t^T \Sigma A^T t) \exp(it^T b - \frac{1}{2} t^T V^T t). \] (5.3)

Equating the imaginary parts of (5.3) yields \( \mu = (I - A)^{-1} b \), which is the MLE \( \hat{\theta} \).

Equating the real parts of (5.3) leads to
\[ \Sigma - A \Sigma A^T = V. \] (5.4)

The formal solution of (5.4) is given by
\[ \Sigma = \sum_{i=1}^{\infty} A^{i-1} V (A^T)^{i-1}, \] (5.5)

if the series converges (Smith, 1967). In fact, this solution is the limit of the iterative equation
\[ \Sigma^{(m+1)} = V + A \Sigma^{(m)} A^T, \]

if the iterations converge. Note that there exist constants \( r \) and \( M \) such that \( ||A|| \cdot ||A^T|| < r < 1 \) and \( ||A^{i-1}|| \cdot ||(A^T)^{i-1}|| \leq Mr^{k-1} \). Therefore, the series in (5.5) converges to a unique solution (Householder, 1964).

To get bounds for \( ||\Sigma|| \), first note if \( \Sigma \) satisfies (5.5), then \( \Sigma^T \) also satisfies (5.5). Since the solution to (5.5) is unique, \( \Sigma \) is symmetric. In addition, \( V \) is positive definite implies that \( \Sigma \) is positive semi-definite because \( A^{i-1} V (A^T)^{i-1} \) are positive semi-definite for all \( i = 1, 2, \ldots \). The matrix \( \Sigma \) is positive definite if \( A \) is of full rank.

Let \( V \) has eigenvalues \( \lambda_i \) and \( V = \Gamma D \Gamma^T \), where \( \Gamma \) is orthogonal and \( D = \text{diag}(\lambda_1, \ldots, \lambda_k) \). Premultiply both sides of (5.5) by \( \Gamma^T \) and postmultiply by \( \Gamma \) gives
\[ \Gamma^T \Sigma \Gamma = \Gamma^T A \Sigma A^T \Gamma = D. \] (5.6)

Denote \( \Gamma^T \Sigma \Gamma \) by \( \Sigma \) and \( \Gamma^T A \Gamma \) by \( B \) and rewrite (5.6) as
\[ \Sigma = B \Sigma B^T + D. \]
Both $B \Sigma B^T$ and $D$ are positive semi-definite matrices. Moreover, $\rho(B \Sigma B^T) = \rho(\Sigma B^T B)$ holds (Marshall and Olkin, 1979, p.216; Mirsky, 1955, p.200). Therefore,

$$
\rho(\Sigma) &= \rho(B \Sigma B^T + D) \\
&\leq \rho(B \Sigma B^T) + \rho(D) \\
&= \rho(\Sigma B^T B) + \|V\|,
$$

which leads immediately to

$$
\rho(\Sigma) - \|V\| \leq \rho(B^T B \Sigma) \leq \rho(B^T B) \rho(\Sigma).
$$

But $\rho(B^T B) = \rho(\Gamma^T A^T \Gamma \Gamma^T A \Gamma) = \rho(A^T A) = \|A\|^2$, and $\rho(\Sigma) = \rho(\Gamma^T \Gamma) = \|\Gamma\|$. Therefore, from (5.7), $\|\Gamma\|(1 - \|A\|^2) \leq \|V\|.$

To obtain a lower bound note that the matrix $\Sigma = \Gamma \Gamma^T$ is positive semi-definite, which implies that $\Sigma - D = B \Sigma B^T \geq 0$. Therefore $\rho(\Sigma) \geq \rho(D) = \|V\|$ (Marshall and Olkin, 1979, p.510). ♦

6 Non-linear Case

We now present two results concerning the stationary distribution of (3.9) where $T$ is non-linear and the noise is dependent on $\theta^{(m)}$. In the first lemma, we assume that the Markov process is geometric ergodic and the second moment of the stationary distribution is finite. Note that these conditions can be checked using Theorem 4 or Theorems 2 and 3.

**Lemma 3** If the Markov process defined by (3.9) is geometric ergodic, its stationary distribution $\Psi$ has finite second moment and $T$ has $\hat{\theta}$ as its stable fixed point, then for $\theta \sim \Psi$,

$$
E\|\theta - \hat{\theta}\|^2 = O(\frac{1}{n}).
$$

**Proof.** Let $T : \mathbb{R}^k \rightarrow \mathbb{R}^k$ has the $k \times k$ matrix $DT(\theta)$ as its first derivative. With an appropriate starting value $\theta = \theta^{(0)}$, we have,

$$
\theta^{(1)} = T(\theta^{(0)}) + \frac{1}{\sqrt{n}} S(\theta^{(0)}) \varepsilon^{(1)},
$$

15
\[ \theta^{(2)} = T(\theta^{(1)}) + \frac{1}{\sqrt{n}} S(\theta^{(1)}) \varepsilon^{(2)} \]
\[ = T \left( T(\theta^{(0)}) + \frac{1}{\sqrt{n}} S(\theta^{(0)}) \varepsilon^{(1)} \right) + \frac{1}{\sqrt{n}} S(\theta^{(1)}) \varepsilon^{(2)} \]
\[ = T^2(\theta^{(0)}) + \frac{1}{\sqrt{n}} \left[ M^{(0)} S(\theta^{(0)}) \varepsilon^{(1)} + S(\theta^{(1)}) \varepsilon^{(2)} \right], \tag{6.1} \]

where \( M^{(0)} = DT(\xi^{(0)}) \), and
\[ \xi^{(0)} = T(\theta^{(0)}) + \frac{1}{\sqrt{n}} \Delta^{(0)} S(\theta^{(0)}) \varepsilon^{(1)}, \tag{6.2} \]

for some \( \Delta^{(0)} = \text{diag}(\delta^{(0)}_i) \) where \( 0 < \delta^{(0)}_i < 1 \). Successive substitution of \( \theta^{(m)} \) into (3.9) and Taylor expanding as in (6.1) gives
\[ \theta^{(m)} = T^m(\theta^{(0)}) + \frac{1}{\sqrt{n}} \left[ M^{(m-2)} \ldots M^{(0)} S(\theta^{(0)}) \varepsilon^{(1)} + M^{(m-2)} \ldots M^{(1)} S(\theta^{(1)}) \varepsilon^{(2)} + \ldots \right. \]
\[ + M^{(m-2)} S(\theta^{(m-2)}) \varepsilon^{(m-1)} + S(\theta^{(m-1)}) \varepsilon^{(m)} \right], \tag{6.3} \]

where \( M^{(j)} = DT(\xi^{(j)}) \), \( j = 0, \ldots, m - 2 \), are defined in a similar way as in (6.2). As \( m \to \infty \), \( T^m(\theta^{(0)}) \to \hat{\theta} \). From (6.3), use the finite second moment assumption gives \( E||\theta - \hat{\theta}||^2 = O(1/n) \). \( \diamond \)

As an extension to the linear case (5.1), the next theorem, Theorem 6 deals with the general case in which the kernel is the stochastic difference equation (3.9)
\[ \theta^{(m+1)} = T(\theta^{(m)}) + \frac{S(\theta^{(m)})}{\sqrt{n}} \varepsilon^{(m+1)}, \]

where \( \varepsilon \sim \mathcal{N}(0, I) \) and \( T \) is not necessarily linear. It is inevitable that stronger conditions have to be imposed. The main assumptions required in Theorem 6 are bounded on \( S \) and the second derivative of the components of \( T \), the ergodicity of the chain and the existence of moments. Also, \( T \) is assumed to be well behaved so that \( T^m(\theta^{(0)}) \to \hat{\theta} \) for some starting value \( \theta^{(0)} \). Theorem 6 applies to cases where the stochastic equation (3.9) holds, regardless of the model. It also gives the asymptotic distribution of the Markov chain \( \{\theta^{(m)}\} \).
Theorem 6 Let the EM operator \( T(\theta) \) in (3.9) be decomposed into \( k \) components such that \( T(\theta) = (T_1(\theta), \ldots, T_k(\theta)) \). For each \( i = 1, \ldots, k \), \( T_i \) is a mapping from \( \mathbb{R}^k \) to \( \mathbb{R} \). Denote the \( ij \)-th element in the matrix \( S(\theta) \) by \( s_{ij}(\theta) \) and the matrix \( DT(\theta) = \frac{\partial T}{\partial \theta_j} \) by \( A(\theta) \), the population value of \( \theta \) by \( \theta^* \). If \( \theta^{(m)} \) converges to the stationary distribution \( \Psi \) with finite sixth moment and the following assumptions hold

- (A.1) \( 0 < ||A(\hat{\theta})|| = \rho < 1 \),
- (A.2) The spectral norm of the second derivative of \( T_i \), \( D^2T_i(\theta) \) is less than \( c_1||\theta|| \) for some constant \( c_1 \),
- (A.3) The spectral norm of the matrix \( Ds_{ij}(\theta)Ds_{ij}^T(\theta) \) is less than \( c_2||\theta||^2 \) for some constant \( c_2 \),

then for \( \theta^{(\infty)} \sim \Psi \),

1. \( E||\theta^{(m)} - \hat{\theta}||^q \) is bounded uniformly by \( O(n^{-q/2}) \) term for \( q = 1, \ldots, 6 \),
2. \( E(\theta^{(\infty)}) = \hat{\theta} + O(1/n) \),
3. \( E||\theta^{(\infty)} - E\theta||^2 = O(1/n) \),
4. as \( n \to \infty \), \( \sqrt{n}(\theta^{(\infty)} - \theta^*) \) tends to \( N(0, K(\theta^*)) \), where

\[
\frac{||S(\theta^*)||^2}{1 - ||A(\theta^*)||^2} = ||K(\theta^*)|| > ||S(\theta^*)||^2.
\]

Proof. To prove (1), observe that in (6.3), \( T^{(m)}(\theta^{(0)}) \) can be made arbitrarily close to \( \hat{\theta} \) by choosing a sufficiently large \( m \). Because the sixth moment of \( \Psi \) is finite, for sufficiently large \( m \), (6.3) guarantees that \( E||\theta^{(m)} - \hat{\theta}||^q \) is bounded uniformly by \( O(n^{-q/2}) \) term for \( q = 1, \ldots, 6 \).

To prove (2), let the \( k \times 1 \) vector \( U^{(m)} = \theta^{(m)} - \hat{\theta} \). For the sake of simplicity in notation, suppress the superscript \( (m) \) when there is no ambiguity. Expectations are always taken conditional on the observed data \( y \) but we suppress the notation of \( y \) for clarity. Subscripts \( i \) on \( U, \theta \) and \( S \) denote the \( i \)-th component of \( U, \theta \), and the \( i \)-th row of \( S \) respectively. Consider the \( i \)-th component in (3.9):

\[
\theta_i^{(m+1)} = T_i(\theta^{(m)}) + \frac{S_i^T(\theta^{(m)})}{\sqrt{n}}e^{(m+1)},
\]

17
and use a Taylor expansion on $T_i : \mathbb{R}^k \to \mathbb{R}$ about $\hat{\theta}$:

$$
\theta_i^{(m+1)} - \hat{\theta}_i = (DT_i)^T U + \frac{1}{2} U^T D^2 T_i(\dot{\xi}_i) U + \frac{1}{\sqrt{n}} S_i^T (\theta^{(m)}) e^{(m+1)},
$$

(6.4)

where $\dot{\xi}_i = \dot{\theta} + \Delta_i U$ for some $\Delta_i = \text{diag}(\delta_{ij})$, $0 < \delta_{ij} < 1$, and $DT_i = \partial T_i / \partial \theta (\hat{\theta})$.

Condition on $\theta^{(m)} = \theta$ and take expectation on both sides of (6.4), we obtain, with a slight abuse of notation,

$$
E[U_i^{(m+1)} | \theta^{(m)}] = (DT_i)^T U + \frac{1}{2} U^T D^2 T_i(\dot{\xi}_i) U,
$$

(6.5)

Further take expectation on both sides of (6.5) with respect to $\theta$, sum over $i$, and let $A = DT(\hat{\theta})$, $X^{(m)} = EU^{(m)}$ and $b^{(m)}$ be a vector whose $i$-th element is $b_i^{(m)} = (1/2) E\{(U^{(m)})^T D^2 T_i U^{(m)}\}$,

$$
X^{(m+1)} = AX^{(m)} + b^{(m)}.
$$

(6.6)

follows. Therefore

$$
X^{(m+1)} = A^{m+1} X^{(0)} + A^m b^{(0)} + \ldots + A b^{(m-1)} + b^{(m)}.
$$

By assumption (A.2), and (1), $D^2 T_i(\xi)$ is bounded by $c_1 ||\xi|| \leq ||\dot{\theta}|| + \delta ||U||$ for some $0 < \delta < 1$, and $E||U||^2$ is $O(1/n)$. Thus $b_i^{(m)}$ is $O(1/n)$. Denote the vector whose components form the bound for $b_i^{(m)}$ by $b$,

$$
||X^{(m+1)}|| \leq ||A^{m+1} X^{(0)}|| + ||A^m b^{(0)} + \ldots + A b^{(m-1)} + b^{(m)}||
\leq ||A^{m+1} X^{(0)}|| + ||A^m b^{(0)}|| + \ldots + ||A b^{(m-1)}|| + ||b^{(m)}||
\leq ||A^{m+1} X^{(0)}|| + ||A||^m ||b^{(0)}|| + \ldots + ||A|| ||b^{(m-1)}|| + ||b^{(m)}||
\leq ||A^{m+1} X^{(0)}|| + (||A||^m + \ldots + ||A||) ||b||
\leq ||A^{m+1} X^{(0)}|| + \frac{1}{1 - ||A||} ||b||.
$$

Since $||A^{m+1} X^{(0)}|| \to 0$ and $||b||$ is of order $O(1/n)$,

$$
\lim_{m \to \infty} EU^{(m+1)} = O(1/n).
$$

(6.7)

Moreover, from the boundedness result in (1), the sequence $\{\theta^{(m)}\}$ is uniformly integrable, this permits the interchange of $\lim_{m \to \infty}$ and expectation in (6.7):

$$
E[\lim_{m \to \infty} \theta^{(m+1)} - \hat{\theta}] = O(1/n),
$$

18
implying result (2).

Result (3) is a direct consequence of the triangle inequality:
$$||\theta^{(\infty)} - E\theta|| \leq ||\hat{\theta}^{(\infty)}|| - ||\hat{\theta}|| + ||\hat{\theta} - E\theta||.$$ 

Squaring both sides gives
$$||\theta^{(\infty)} - E\theta||^2 \leq ||\theta^{(\infty)}||^2 - ||\hat{\theta}||^2 + ||\hat{\theta} - E\theta||^2 + 2||\theta^{(\infty)} - \hat{\theta}|| ||\hat{\theta} - E\theta||$$

Since $E||\theta^{(\infty)} - \hat{\theta}||^2 = O(1/n)$ by Theorem 3 and $||\hat{\theta} - E\theta||^2 = O(n^{-2})$, taking expectation on both sides gives the result $E||\theta^{(\infty)} - E\theta||^2 = O(1/n)$.

Asymptotic normality can be proved by mimicking the linear case. We treat $n$ as fixed and linearize $T$ and $S$. First use Taylor expansion on each component $s_{ij}$ of $S_i(\theta)$ in (6.4).

$$S_i^T(\theta)e^{(m+1)} = \sum_{j=1}^{k} s_{ij}(\theta)e_j^{(m+1)} = \sum_{j=1}^{k} \{s_{ij}(\hat{\theta}) + (Ds_{ij}(\xi_{ij}))^TU\}e_j^{(m+1)}.$$  

(6.8)

where $e_j^{(m+1)}$ is the j-th component of $e^{(m+1)}$ and $\xi_{ij} = \hat{\theta} + T_{ij}U$, $T_{ij}$ being a diagonal matrix with elements all lie between 0 and 1. Note that under assumption (A.3),

$$||Ds_{ij}(\xi_{ij})^TU|| < c_2||\hat{\theta} + T_{ij}U||^2 
\leq \gamma_0 + \gamma_1||U|| + \gamma_2||U||^2$$ for some $\gamma_0, \gamma_1, \gamma_2.$

(6.9)

Substituting (6.8) in (6.4) gives

$$\theta_i^{(m+1)} - \hat{\theta}_i = (DT_i)^TU + \frac{1}{2}U^TD^2T_i(\xi_{ij})U + \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \{s_{ij}(\hat{\theta}) + (Ds_{ij}(\xi_{ij}))^TU\}e_j^{(m+1)}.$$  

(6.10)

Multiply (6.10) throughout by $\sqrt{n}$, denote $\sqrt{n}(\theta_i^{(m+1)} - \hat{\theta}_i)$ by $W_i^{(m+1)}$ and still suppressing the superscript $m$ in $U^{(m)}$:

$$W_i^{(m+1)} = (DT_i)^TW_i^{(m)} + \frac{\sqrt{n}}{2}U^TD^2T_iU + \sum_{j=1}^{k} s_{ij}(\hat{\theta})e_j^{(m+1)} + \sum_{j=1}^{k} (Ds_{ij}^TU)e_j^{(m+1)}.$$  

(6.11)
Identify a new linear order one autoregressive process \( \{Z^{(m)}\} \) by defining
\[
Z_i^{(m+1)} = (DT_i)^T Z_i^{(m)} + \sum_{j=1}^{k} s_{ij}^T \hat{\theta} \varepsilon_j^{(m+1)},
\]
with \( Z_i^{(0)} = W_i^{(0)} \). Expression (6.11) becomes
\[
W_i^{(m+1)} - Z_i^{(m+1)} = (DT_i)^T (W_i^{(m)} - Z_i^{(m)}) + \frac{\sqrt{n}}{2} U^T D^2 T_i U + \sum_{j=1}^{k} (D s_{ij}^T U) \varepsilon_j^{(m+1)}.
\] (6.13)

First we show that the process \( E R^{(m)} = E(W^{(m)} - Z^{(m)}) \) is bounded by an order \( O(1/n) \) term.

Put together the components \( Z_i^{(m)} \) in (6.12) to form the linear process \( \{Z^{(m)}\} \). In vector form,
\[
Z^{(m+1)} = A(\hat{\theta}) Z^{(m)} + S(\hat{\theta}) \varepsilon^{(m+1)},
\]
Also form \( W^{(m)} = \sqrt{n}(\theta^{(m)} - \hat{\theta}) \) by putting \( W_i^{(m)} \) together.

Taking expectation with respect to \( \theta^{(m)} = \theta \), we obtain
\[
E[R^{(m+1)} | \theta^{(m)}] = (DT) E[R^{(m)} | \theta^{(m)} = \theta] + \frac{\sqrt{n}}{2} P^{(m)},
\]
where \( P^{(m)} \) is a vector with the i-th component equals \( U^T D^2 T_i U \). Further take expectation with respect to \( \theta \) and by result in (1),
\[
ER^{(m+1)} = (DT)(ER^{(m)}) + O(1/n),
\]
follows. This recursive equation is analogous to (6.6) and thus \( ER^{(m)} \) is bounded by an order \( 1/n \) term.

Our next goal is to prove that \( E||W^{(m)} - Z^{(m)}||^2 = O(1/n) \) as \( m \to \infty \). Squaring both sides of (6.13) gives
\[
(W_i^{(m+1)} - Z_i^{(m+1)})^2 = [(DT_i)^T (W_i^{(m)} - Z_i^{(m)})]^2 + \frac{n}{4} [U^T D^2 T_i U]^2
\]
\[
+ \frac{\sqrt{n}}{2} (DT_i)^T (W_i^{(m)} - Z_i^{(m)}) [U^T D^2 T_i U]
\]
\[
+ (\sum_{j=1}^{k} (D s_{ij}^T U) \varepsilon_j^{(m+1)})^2
\]
+ cross terms involving \( \varepsilon^{(m+1)} \). (6.14)
We first sum over \(i\), then take expectation conditional on \(\theta^{(m)} = \theta\) and then uncondition by further taking expectation with respect to \(\theta\) conditional on \(y\). When we sum (6.14) across \(i = 1, \ldots, k\), the first term in the right hand side of (6.14) is exactly \(\|DTR^{(m)}\|^2\), which is less than \(\rho^2 \|R^{(m)}\|^2\) where \(\rho < 1\) is the spectral norm of \(DT\). The second term is bounded by a constant times \(n\|U\|^6\). For the third term, use the proven fact that \(E(W^{(m)} - Z^{(m)})\) is bounded by \(O(1/n)\) term. By virtue of (A.2) the third term is bounded by \(O(1/n)\) term. The square of the fourth term, when summed across \(i\) is

\[
\sum_{i=1}^{k} \sum_{j,p} \varepsilon_j^{(m+1)} \varepsilon_p^{(m+1)} U^T Ds_{ij} DT s_{ip} U. \tag{6.15}
\]

Take expectation conditional on \(\theta^{(m)} = \theta\), all the terms with \(j \neq p\) vanish because \(E\varepsilon_j^{(m+1)} \varepsilon_p^{(m+1)} = 0\). From 6.9, \(U^T Ds_{ij} DT s_{ij} U = \|DT s_{ij} U\|^2\) is essentially \(O(1/n)\). Thus, the sum in (6.15) is still \(O(1/n)\).

Summing up, the following relation is now established:

\[
E\|W^{(m+1)} - Z^{(m+1)}\|^2 \leq \rho^2 E\|W^{(m)} - Z^{(m)}\|^2 + \frac{c}{n},
\]

for some constant \(c\).

Therefore the sequence \(\{E\|W^{(m)} - Z^{(m)}\|^2\}\) converges to a term of order \(1/n\) as \(m \to \infty\). To indicate that \(W^{(m)}\) and \(Z^{(m)}\) actually depend on \(n\), we write

\[
\lim_{m \to \infty} E\|W^{(m)}_{(n)} - Z^{(m)}_{(n)}\|^2 = O(1/n). \tag{6.16}
\]

We still require the boundedness of the fourth moment in order to interchange the expectation and limit in (6.16).

When \(m\) is sufficiently large, \(E\|W^{(m)}\|^4\) is bounded using the finite fourth moment assumption. The expectation \(E\|Z^{(m)}\|^4\) is also bounded because \(Z^{(m)}\) is a linear autoregressive process that converges to a stationary distribution which is Gaussian. Therefore its fourth moment is also finite. As a result

\[
E\|W^{(m)} - Z^{(m)}\|^4 \leq E\|Z^{(m)}\|^4 + E\|W^{(m)}\|^4 + 2E(\|Z^{(m)}\|^2 \|W^{(m)}\|^2)
\]

\[
E\|Z^{(m)}\|^4 + E\|W^{(m)}\|^4 + 2(\|Z^{(m)}\|^4 E\|W^{(m)}\|^2)^{1/2}.
\]

21
Hence, $E||W^{(m)} - Z^{(m)}||^4$ is bounded uniformly for sufficiently large $m$. This allows the interchange of $\lim_{m \to \infty}$ and expectation in (6.16):

$$E||W^{(\infty)}_{(n)} - Z^{(\infty)}_{(n)}||^2 = O(1/n), \quad (6.17)$$

where $W^{(\infty)}_{(n)} = \lim_{m \to \infty} W^{(m)}_{(n)}$ and $Z^{(\infty)}_{(n)} = \lim_{m \to \infty} Z^{(m)}_{(n)}$.

Theorem 5 asserts that the stationary distribution of $\{Z^{(m)}\}$ is given by

$$Z^{(\infty)}_{(n)} \sim \mathcal{N}(0, K(\hat{\theta})),$$ \quad (6.18)

putting the subscript $(n)$ back. The variance matrix $K(\hat{\theta})$ equals to $\sum_{i=1}^{\infty} A_i S S^T A_i^{-1}$ with $A, S$ evaluated at $\hat{\theta}$. Also, from Theorem 5, $||S(\hat{\theta})||^2/(1 - ||A||^2) \geq ||K(\hat{\theta})|| > ||S(\hat{\theta})||^2$. Under continuity assumptions on $S$ and $T$,

$$\lim_{n \to \infty} Z^{(\infty)}_{(n)} = Z^{(\infty)}_{(\infty)},$$

where $Z^{(\infty)}_{(\infty)} \sim \mathcal{N}(0, K(\theta^*))$, using the consistency property of the MLE $\hat{\theta}$.

From (6.17),

$$E||W^{(\infty)}_{(n)} - Z^{(\infty)}_{(n)}||^2 \to 0 \quad \text{as} \quad n \to \infty.$$ \quad \text{The mean square convergence implies}

$$W^{(\infty)}_{(n)} - Z^{(\infty)}_{(n)} \to 0 \quad \text{in probability.}$$

By Slutsky's theorem, $Z^{(\infty)}_{(n)} \to Z^{(\infty)}_{(\infty)}$ in law implies $W^{(\infty)}_{(n)} \to Z^{(\infty)}_{(\infty)}$ in law. Therefore, for $\theta^{(\infty)} \sim \Psi_{n}$,

$$\sqrt{n}(\theta^{(\infty)} - \theta^*) \to \mathcal{N}(0, K(\theta^*)).$$ \quad (6.19)

\begin{remark}

1. Normality assumption on $\epsilon^{(m)}$ in (3.9) is not required in proving part (1),(2) and (3).

2. For Examples 2.1 and 3.7, all of the conditions (A.1)-(A.3) are satisfied and Theorem 6 applies.

\end{remark}
Theorem 1 states that when the complete data are generated from an exponential family, the Markov kernel of the Stochastic EM iterates is approximately Gaussian and allows the decomposition,

$$\theta^{(m+1)} = T(\theta^{(m)}) + \epsilon^{(m+1)},$$

where $\epsilon^{(m+1)} \sim \mathcal{N}(0, V(\theta^{(m)}))$, $V(\theta) = DT(\theta)J^{-1}_{\text{cond}}(\theta)(DT(\theta))^T$. On the other hand, Theorem 6 asserts optimal properties of the stationary distribution. The two results can be combined using Lemma 2. The following corollary reveals an approximate relation between the variance of the stationary distribution and the information matrix, thereby making our results more interpretable. Continue with the above notation and suppose that the conditions of Theorem 6 hold, then

**Corollary 3** \( \theta^{(\infty)} - \theta^* \rightarrow \mathcal{N}(0, Q(\theta^*)) \) where

$$\frac{||J_{\text{cond}}J^{-2}_{e}||}{1 - \rho^2} \geq Q(\theta^*) \geq ||J_{\text{cond}}J^{-2}_{e}||,$$

with the information matrices all evaluated at $\theta^*$.

**Proof.** We only need to prove (3). Equation (6.18) in Theorem 6 implies

$$\frac{n||V(\hat{\theta})||}{1 - \rho^2} \geq ||K(\hat{\theta})|| \geq n||V(\hat{\theta})||.$$

By Lemma 2, $V(\hat{\theta}) = J_{\text{cond}}J^{-1}_{e}J^{-1}_{\text{cond}}J^{-1}_{e}$ with $J_{\text{cond}} = J_{\text{cond}}(\hat{\theta})$ and $J^{-1}_{e} = J^{-1}_{e}(\hat{\theta})$. Using the fact that for any square matrices $A$ and $B$, $||BA|| = ||AB|| \leq ||A|| ||B||$ and following the proof in Theorem 6, result (3) is established. \( \diamond \)

**Remark:** The variance term $Q(\theta^*)$ is of order $1/n$. To make the result interpretable, we do not multiply $(\theta^{(\infty)} - \theta^*)$ by $\sqrt{n}$ in order to avoid the awkward looking $n$ term in the variance.
References


