A STOCHASTIC EM ESTIMATOR
IN THE PRESENCE OF MISSING DATA —
THEORY AND APPLICATIONS

by

E.H.S. Ip

Technical Report No. 304
November 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 EM ESTIMATOR
IN THE PRESENCE OF MISSING DATA —
THEORY AND APPLICATIONS

by

Eddie H.S. Ip

Technical Report No. 304
November 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 EM estimator in the presence of missing data – theory and applications

Eddie Hak-sing Ip
November 1994

Summary

This thesis provides a study of a Monte Carlo version of the EM (for Expectation-Maximization) algorithm for handling complex missing-data structure in which high-dimensional integrations may be involved. Assuming a parametric model for the complete data, we propose a method for imputing values for missing data and then iteratively perform direct parametric inference based on the pseudo-complete data. If the iteration converges, the result of the procedure is a sample from a stationary distribution derived from the Markov chain formed by the iterations of the parameter. This algorithm is called Stochastic EM and the estimator we propose is the mean of the stationary distribution.

We provide new theoretical results on the Stochastic EM estimator. First, under an exponential family setting, we show that asymptotically the Markov chain kernel of Stochastic EM can be decomposed into an additive stochastic difference equation. This decomposition allows us to provide useful results for the asymptotic behavior of the stationary distribution under specific conditions. The next result we obtain is that the Stochastic EM estimator is close to the maximum likelihood estimate. Another result shows that asymptotically the Stochastic EM iterations converge to a normal distribution with the true parameter as its mean. Bounds are obtained for the norm of the covariance matrix.

There are two examples, one from medical science and one from education to substantiate the theory. In the education example, the use of straightforward EM would require performing an overwhelmingly large number of high-dimensional numerical integrations even for a moderate sample of a thousand multivariate binary observations. In contrast, Stochastic EM avoids numerical integration altogether and relies on simulation techniques to obtain an approximate solution. We exploits recent simulation technologies such as the Gibbs sampler as a means to reconstruct the missing data. Another contribution of this thesis to statistical methodology is our new method that computes the standard error of the estimate without having to write new codes. Finally some of potential applications of the Stochastic EM algorithm to problems such as pharmacokinetic modeling where non-parametric functions are present in a hierarchical model are discussed.
Contents

1 Introduction 4

2 Imputation Methods for Missing Data 10
  2.1 Overview 11
  2.2 Missing Information Principle and EM 14
  2.3 The EM Algorithm 17
  2.4 Limitations of EM 19
  2.5 Filling in Stochastically 21

3 Stochastic EM 24
  3.1 Stochastic EM 24
  3.2 Ergodicity of Stochastic EM 29
    3.2.1 Markov process preliminary 29
    3.2.2 Ergodicity Under Compactness 30
    3.2.3 Ergodicity under general conditions 33
  3.3 Examples of Stationary Distribution 35
  3.4 Grouped Multinomial Example 44

4 Stationarity in Exponential Families 47
  4.1 Preliminary 48
    4.1.1 Exponential families 48
    4.1.2 Information 50
  4.2 Mean of the Stationary Distribution 51
  4.3 Decomposition of Stochastic Difference Equation 56
  4.4 Ergodic Result and Existence of Moments 61
  4.5 Linear Autoregressive Process 62
  4.6 Non-linear Autoregressive Process 67
5 Binary data analysis
  5.1 Introduction ........................................... 78
    5.1.1 Latent variable ................................... 78
    5.1.2 Gibbs Sampler ................................... 80
  5.2 Variance Component .................................. 81
  5.3 Empirical Bayes with SAT Data ......................... 84
    5.3.1 Background and issues .......................... 84
    5.3.2 Model and inference ............................. 87
    5.3.3 Results in analyzing SAT data ................ 90

6 Standard Error ......................................... 93
  6.1 Monte Carlo Method .................................. 93
  6.2 Censored Weibull Data ................................ 94
  6.3 Genetic Linkage ..................................... 99

7 Concluding Remarks .................................... 102
Chapter 1

Introduction

This study deals with the methodological issue of how to perform statistical inference in missing data problems when conventional methods are infeasible. In this introductory chapter, we consider three missing data examples as motivation for the method we propose.

Example 1.1. Medical data

This example illustrates how missing data problem could arise in a non-trivial situation. Data are not in a real sense missing but the setup of the model allows ‘missing data’ statistical methods to come into play.

The data set is originally reported by Weil (1970) and later analyzed by Ochi and Prentice (1984) and McCulloch (1994).

A clinical trial was performed on pregnant rats to test the effect of a chemically treated diet. In the experiment, 16 pregnant rats were given a chemically treated diet while 16 were given a control diet. In each litter, offspring alive after four days were followed to ascertain their survival status at 21 days.

For each rat, the outcome $Y$, whether it survives or not, is binary with

$$Y = \begin{cases} 
1, & \text{if the rat survives,} \\
0, & \text{otherwise.}
\end{cases}$$

We use $i = 1, 2$, to index the groups, $j = 1, \ldots, 16$, to index the litters and $k = 1, \ldots, n_j$ to index the individuals. Ochi and Prentice (1984) model the probability of survival of the $k$-th rat in the $j$-th litter within the $i$-th group as a function of the group effect, the litter effect and a random variation at
an individual level. Specifically,

\[ \text{Prob}(Y_{ijk} = 1 \mid \mu_i, \beta_{ij}) = \Phi(\mu_i + \beta_{ij}). \]

where \( \Phi \) is the standard normal cumulative density. We set up the example as a missing data problem. Introduce a latent variable \( Z_{ijk} \) such that

\[ Z_{ijk} = \mu_i + \beta_{ij} + \varepsilon_{ijk}, \]

where \( \varepsilon_{ijk} \sim \mathcal{N}(0, 1) \), and

\[ Y_{ijk} = I(Z_{ijk} > 0). \] (1.1)

where \( I(.) \) is the indicator function. Here, \( \mu_i \) is the treatment effect on the latent scale and \( \beta_{ij} \) is the litter effect. Further, the litter effect is regarded as random sample from a larger population:

\[ \beta_{ij} \sim \mathcal{N}(0, \sigma_i^2), \]

independently. Although in reality no raw data was lost in the original experiment, the problem can be formulated as having a missing-data structure by treating the variables \( Z_{ijk} \) and parameters \( \beta_{ij} \) as missing. The statistical structure of this example is not complex. Conventional optimization method can be used to solve the problem. However, there are a great deal of similar 'missing data' problems in the medical literature that conventional optimization method cannot easily handle. One example is the study of mixed models for ordered categorical data. Harville and Mee (1984) remark that the evaluation of the maximum likelihood estimate (MLE) in these cases “will present an insurmountable computational problem, calling, for example, for the integration of pdf \( \phi_N(0, V) \) over an \( N \)-dimensional rectangular region.” Here \( \phi_N \) is the \( N \)-variate Gaussian density function, where \( N \) is the number of observable data points.

The problem of numerical integration is especially severe when the statistical model becomes highly multivariate in nature. The following simple example demonstrates why inference may be difficult in these situations.

**Example 1.2. Time series**

Consider the following time series example.

\[ X_t = \theta X_{t-1} + \varepsilon_t, \quad t = 1, \ldots, n, \]
where $0 < \theta < 1$, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ independently, and $X_0$ is generated from the stationary distribution, which is $\mathcal{N}(0, \sigma^2/(1 - \theta^2))$. The value of $X_0$ is given.

Assume $\sigma^2$ to be known with the goal being to estimate $\theta$. With complete data available, the solution is straightforward. The MLE $\hat{\theta}$, for example, is

$$\hat{\theta} = \frac{\sum_{t=1}^{n} x_t x_{t-1}}{\sum_{t=1}^{n} x_t^2}.$$  

Now suppose that the data are censored so that we only observe those $X_t$ greater than their predecessors, i.e.,

$$Y_t = \begin{cases} X_t, & \text{if } X_t > X_{t-1}, \\ \ast, & \text{if } X_t \leq X_{t-1}, \end{cases} \quad (1.2)$$

where $Y_t$ is the observed data and $\ast$ represents a missing datum. With such a censoring mechanism, the originally simple problem suddenly becomes difficult to solve.

We can employ conventional statistical methodology such as the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977) to find the MLE. The EM algorithm is an iterative scheme. At the $m$-th iteration we calculate the conditional expectations of the sufficient statistics $\sum_{t=1}^{n} x_t x_{t-1}$ and $\sum_{t=1}^{n} x_t^2$, of the complete data, given the observed data $\{Y_t\}$, and the current estimate $\theta^{(m)}$ of $\theta$. The conditional expectations take the form of integrals over restricted orthants in high-dimensional space. As an example, consider the following simple censored time series $(y_1, \ast, \ast, y_4)$ generated from (1.2).

The conditional expectations of the sufficient statistics are

$$\int \int (x_0 x_1 + x_1 x_2 + x_2 x_3 + x_3 x_4) \phi(x_2, x_3 \mid x_1 = y_1, x_4 = y_4, \theta = \theta^{(m)}) \, dx_2 dx_3,$$  

(1.3)

$$\int \int (x_1^2 + x_2^2 + x_3^2 + x_4^2) \phi(x_2, x_3 \mid x_1 = y_1, x_4 = y_4, \theta = \theta^{(m)}) \, dx_2 dx_3,$$  

(1.4)

where the integrals are evaluated over the space $\{(x_2, x_3) : y_1 > x_2 > x_3, x_3 < y_4\}$. The conditional density $\phi(x_2, x_3 \mid x_1 = y_1, x_4 = y_4, \theta = \theta^{(m)})$ is bivariate normal. However, when the chain of missing data becomes longer, integrations of the form (1.3) and (1.4) become much more difficult to compute.

In order to avoid numerical integrations of the type (1.3), we propose and investigate a stochastic version of the EM algorithm called Stochastic EM
(Celeux and Diebolt 1985) that utilizes current Monte Carlo technologies such as the Gibbs sampler. Suppose that the complete data are generated from a family of distributions parameterized by \( \theta \). In order to estimate \( \theta \), the Stochastic EM algorithm imputes values for the missing data and estimates \( \theta \) as if the complete data were available. Specifically, beginning with an initial guess \( \theta^{(0)} \) of the parameter, the Stochastic EM algorithm consists of two steps.

1. S-step: draw Monte Carlo samples from the conditional distribution of the missing data given the observed data and \( \theta^{(0)} \), and 'fill in' for the missing data to form a pseudo-complete sample.

2. M-step: find the MLE \( \theta^{(1)} \) based on the pseudo-complete data, and update \( \theta^{(0)} \) to \( \theta^{(1)} \).

Steps (1) and (2) are then performed iteratively until \( \{\theta^{(m)}\} \) attains stationarity. The mean of the stationary distribution is used as an estimator for \( \theta \).

If we apply the Stochastic EM algorithm to the above censored time series example using the Gibbs sampler approach, then it is only necessary to simulate from univariate normal distributions under constraints in the S-step. To be specific, draw Monte Carlo samples from the following conditional distributions:

\[
x_2 \mid y_1, x_3, y_4 \text{ under the constraint } y_1 > x_2 > x_3, \quad (1.5)
\]

and

\[
x_3 \mid y_1, x_2, y_4 \text{ under the constraint } x_2 > x_3 \text{ and } x_3 < x_4. \quad (1.6)
\]

Even with constraints, the Gibbs sampler (1.5) and (1.6) are easier to handle than the integration of (1.3). This is especially true when the dimensionality of the problem is high (Casella and George 1992). See also Gelfand, Smith and Lee (1992) for a discussion on the use of the Gibbs sampler in constrained parameter and truncated data problems under a Bayesian framework.

Example 1.3. Educational data

In a typical standardized educational or psychological test, students/subjects are scored on \( n \) items (questions). The i-th student’s response to the j-th item \( Y_{ij} \), equals 1 when he answers correctly and 0 otherwise. Suppose for a test, \( p \) features of cognitive characteristics are identified. On a mathematics test for example, geometric skill and reading comprehension of word problem could be two of the \( p \) features. For each item the presence of a particular
feature is coded as 1 and 0 otherwise. This binary vector associated with the j-th item is denoted by \( x_j \).

Further assume that each student is characterized by a \( p \times 1 \) ability parameter vector \( \beta_i \) that signifies his or her proficiency in each of \( p \) attributes. Consider the following probability model:

\[
\text{Prob}(Y_{ij} = 1 \mid \beta_i) = \Phi(x_j^T \beta_i),
\]

where \( x_j = (x_{j1}, \ldots, x_{jp})^T \) is a vector of known binary coded covariates for the j-th item. Similar to the hierarchical model in Example 1.1, the ability is regarded as a random effect such that

\[
\beta_i \sim \mathcal{N}_p(\mu, \Sigma).
\]

Denote the parameter of interest \((\mu, \Sigma)\) by \( \theta \). For such a model, finding the MLE requires solving the following function for its maximum with respect to \( \theta \).

\[
\sum\log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_j \{\Phi(\beta_i^T x_j)\}^{u_{ij}} \{1 - \Phi(\beta_i^T x_j)\}^{1-u_{ij}} \phi_p(\beta_{ij}; \theta) \, d\beta_{i1} \cdots d\beta_{ip}.
\]

In order to find the maximum, the EM algorithm evaluates in each of its E-step the expectation of the sufficient statistics \( E(\sum_i \beta_i \mid y, \theta^{(m)}) \) and \( E(\sum_i \beta_i \beta_i^T \mid y, \theta^{(m)}) \). For a particular vector \( \beta_i = (\beta_{i1}, \ldots, \beta_{ip}) \), the E-step involves integrals of the following kind:

\[
E(\beta_{ik} \mid y, \theta^{(m)}) = \frac{\int \cdots \int \beta_{ik} f_i(y) \phi_p(\beta_{ij}; \theta^{(m)}) \, d\beta_{i1} \cdots d\beta_{ip}}{\int \cdots \int f_i(y) \phi_p(\beta_{ij}; \theta^{(m)}) \, d\beta_{i1} \cdots d\beta_{ip}}
\]

where \( f_i(y) = \prod_j \{\Phi(\beta_i^T x_j)\}^{u_{ij}} \{1 - \Phi(\beta_i^T x_j)\}^{1-u_{ij}} \).

In Chapter 5 we discuss a real data set taken from a Scholastic Aptitude Test (SAT) where \( p = 14 \). A quick calculation shows that if the number of students is in the order of 1,000, the EM algorithm easily commands the numerical integration of 10 million 14-dimensional integrals. Surprising as it may sound, the Stochastic EM is able to give an approximate solution to the above maximization problem without having to perform a single numerical integration!

The above examples motivate the consideration of the Stochastic EM methodology both in its theory and applications. The objective of this thesis is two-fold. First, it presents new results concerning the behavior of
the Stochastic EM iterations, in particular, the optimality properties of the mean of the stationary distribution as an estimator. Second, it contributes to methodology by providing applications of the Stochastic EM method to missing data problems where conventional methods cannot be conveniently implemented.

Chapter 2 provides a history of the idea of imputation for the missing data in the statistics literature. The Missing Information Principle and the EM algorithm are briefly described. Chapter 3 deals with the basic concepts of the Stochastic EM methodology. The concepts are illustrated with a number of examples. Most examples are special cases where the stationary distribution of the Stochastic EM iterations can be analytically derived. In general, the stationary distribution of the Stochastic EM iterations is difficult to obtain explicitly. Chapter 4 is devoted to the study of the more general situation. For example, properties of the stationary distribution of the Stochastic EM iterations when the complete data is generated from exponential families are studied. We show that the mean of the stationary distribution, under reasonably regular conditions, is close to the MLE. The asymptotic normality of the stationary distributions for specific situations is established as well. Finally, as a demonstration of the versatility of the Stochastic EM approach to solve missing data problems, we examine two real data examples concerning binary data in Chapter 5. Chapter 6 is devoted to a method of obtaining the standard error of the Stochastic EM estimates via Monte Carlo. Two examples are used to substantiate this idea. We conclude the thesis in Chapter 7 with a discussion of the potential applications of the Stochastic EM algorithm and related research.
Chapter 2

Imputation Methods for Missing Data

The problem of missing data or more generally, missing information has intrigued statisticians for many years. Early efforts to deal with missing data date back to Yates (1930), Wilks (1932) and perhaps Fisher (Healy and Westmascott, 1956). Missing data refers to the case where certain components of a data point are missing. The missing data could be a non-response in a survey or a drop-out patient in a clinical trial. In these situations, data are missing in the data collection process. Little and Rubin (1978) make the distinction between data which are missing-at-random and not missing-at-random. Data which are missing-at-random are easier to handle and a major part of the literature is devoted to methodology for handling such data. Missing information is a more general concept. It includes cases where data are aggregated, censored or truncated. It also includes cases where parameters are modeled as random variables. Latent variables, a concept very common in the psychometric literature, is yet another example where the data is not in a real sense missing. We could regard the latent variables in these models as structural missing data when applying a particular algorithm that handles missing data. Computationally, it is not necessary to distinguish between missing data and missing information.

Because the focus of this thesis is primarily on the computational aspects, we use the terms missing data and missing information interchangeably.
2.1 Overview

Historically, the two kinds of missing data are not studied in a unified way. A large part of the early literature on handling missing data dealt with situations arising from the data collection process out in the field. Many methodologies were first developed for regression and analysis of variance (ANOVA). Early review papers of these methodologies include Afifi and Elashoff (1966), Hartley and Hocking (1971).

Consider a two-way ANOVA table of $T$ treatments and $B$ blocks. Suppose that data in certain cells are missing. The original ANOVA model, which is relatively simple, is suddenly changed into a complex one.

There are some trivial methods to handle such missing data. Obviously the best way to treat missing data problem is not to have them. Thus one way to generate a complete sample is to discard part of the data so that the rest is complete. However, it may happen that nearly all the data points have some missing component or that the missing data have a systematic pattern. The discarded partial data could provide useful information.

In pharmacokinetics experiments, for example, data may be more complete for young and healthy volunteers, whereas observations are scanty for sick and elderly patients, because the tolerance for taking repeated measurements in this population is lower. Discarding incomplete data could lead to a heavy bias in the subsequent analysis when eventually the target population of the experiment is that of the elderly and the sick.

To use all available data, one approach is to ‘fill in’ for the missing data and then proceed as if the data were all present. We hope that when we ‘fill in’ values that are proper in some sense, we might be able to estimate the parameters correctly.

Various methods have been proposed on how to substitute values for the missing data. In the two-way ANOVA case, one idea is to substitute a predicted value for the missing value using the row and column averages. These techniques are known as missing plot techniques. Allan and Wishart (1930) propose using the formula $(T y_{+}^{(1)} + B y_{+}^{(2)} - y_{+})/[(T - 1)(B - 1)]$ to ‘fill in’ when there is only one missing value. The sum of cells with the same treatment and the sum of cells in the same block as the missing data are denoted by $y_{+}^{(1)}$ and $y_{+}^{(2)}$ respectively and $y_{+}$ is the sum of all observed cells. Yates (1933) and Wilkinson (1958) extend this work. Yates performs a least squares analysis and obtains an unbiased estimate. Snedecor (1946) outlines
an iterative procedure when there are more than one missing cell and makes appropriate adjustment on the degrees of freedom in the testing of hypothesis. Bartlett (1937), Hartley (1956) suggest different non-iterative techniques to substitute for the missing data. Healy and Westmascott (1956) describe an iterative scheme similar to the one outlined in Snedecor (1946). The method is to use trial values for the missing data and obtain their predicted values. Complete data analysis is performed and the procedure is iterated until the missing data do not change appreciably. This iterative scheme is one of the earliest application of the EM idea. In fact, it can be shown that this method is actually a special case of the self-consistency principle proposed by Efron (1967) and is equivalent to the EM algorithm.

An early effort to fill in for the missing data with a correct 'guess' and relate it to maximum likelihood estimate (MLE) of the parameter of interest is due to Hartley (1958). He shows for certain discrete distributions, replacing the missing values by their expected values given the model parameters and the observed data would lead to the MLE. Since the model parameters are not known, it implies that an iterative procedure has to be used. In the same paper, he indicates that the method is akin to the missing plot techniques in ANOVA. The examples given in the Hartley paper, though, include only special univariate cases where the score functions are linear in the observations. It turns out that only in these cases, 'fill in' and iterate using the above scheme leads to the MLE.

Federspiel, Monroe and Greenberg (1959) apply a similar procedure in a more complex situation. The iterative method they use again replaces the missing components by their conditional expectations, given the observed components and the model parameters. However, the estimates they obtain are biased. The following example shows why such a procedure could lead to a biased estimate. In this example, the missing pattern is systematic and is not a result of deficient data collection.

**Example 2.1. Unknown Variance.**

Consider $n$ independent and identically distributed random variables $\beta_1, \ldots, \beta_n$, having a common $\mathcal{N}(0, \sigma^2)$ distribution. We observe $y = (y_1, \ldots, y_n)$, where $y_i = \beta_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_i$ are independent of $\beta_i, i = 1, \ldots, n$. Suppose that $\tau^2$ is known and we want to estimate $\theta = \sigma^2$. 

12
Normal theory tells us that the MLE $\hat{\theta}_{ML}$ of the parameter $\theta$ is

$$\hat{\theta}_{ML} = \begin{cases} s^2 - \tau^2, & \text{if } s^2 > \tau^2, \\ 0, & \text{if } s^2 \leq \tau^2, \end{cases}$$

where $s^2 = \frac{1}{n} \sum_{i=1}^{n} y_i^2$. We can regard the complete data as $x_i = (\beta_i, y_i)$, $i = 1, \ldots, n$, where only $y_i$'s are observed and the $\beta_i$'s as missing and proceed with an iterative scheme using Hartley's 'filling in' scheme.

At the m-th iteration, let $\theta^{(m)}$ be a 'guess' of the parameter $\theta = \sigma^2$. Then

$$\beta_i | y_i, \theta^{(m)} \sim \mathcal{N}\left( \frac{\theta^{(m)}}{\theta^{(m)} + \tau^2 y_i}, \frac{\theta^{(m)}}{\theta^{(m)} + \tau^2} \right). \quad (2.1)$$

For each missing $\beta_i$, insert the mean of the conditional distribution, $(\theta^{(m)} y_i)/(\theta^{(m)} + \tau^2)$, and compute the MLE

$$\theta^{(m+1)} = \frac{\theta^{(m)}^2}{(\theta^{(m)} + \tau^2)^2 s^2} \quad (2.2)$$

of the parameter $\theta$ based on the 'filled in' data.

*Figure 2.1: (a) Case: $s^2 > 4\tau^2$; (b) Case: $s^2 \leq 4\tau^2$.*

The expression (2.2) is a first order difference equation of the form $\theta^{(m+1)} = y = f(\theta^{(m)})$. The functions $y = f(\theta)$ and $y = \theta$ are sketched in Figure (2.1) for which there are two possible cases. In the first case, when $s^2 > 4\tau^2$, the
graphs of \( y = f(\theta) \) and \( y = \theta \) intersect at two points and \( \theta^{(m)} \) converges to one real root of the equation

\[
(\theta + \tau^2)^2 = \theta s^2. \tag{2.3}
\]

The legitimate solution of \( \theta \) of the quadratic equation (2.3) is given by

\[
(s^2/2) - \tau^2 + \frac{s^2}{2} \sqrt{1 - 4\tau^2/s^2}.
\]

In the second case, when \( s^2 \leq 4\tau^2 \), the graphs of \( y = f(\theta) \) and \( y = \theta \) only intersect at the origin and \( \theta^{(m)} \) converges to 0. (See Figure 2.1 (b)).

Compared with \( \hat{\theta}_{ML} \), this estimate is biased. When \( \tau^2 \) is small relative to \( s^2 \), a Taylor expansion yields

\[
\sqrt{1 - \frac{4\tau^2}{s^2}} \approx 1 - \frac{2\tau^2}{s^2},
\]

so that the bias is approximately \(-\tau^2\). ♦

Example 2.1 shows that even with the use of a reasonable value, namely the expected value of the conditional distribution given the observed data, to ‘fill in’ and iterate, we obtain a biased estimate. This phenomenon was recognized early on when iterative schemes of this kind were used. Buck (1960) uses a similar approach and tries to correct for the bias in the case of a single component. He, however, fails to give the correct extension to more than a single missing component. Other methods also have been proposed. Nevertheless, as commented by Orchard and Woodbury (1972), “one can determine from their extensive analysis that most approximations are best ‘quick and dirty’ and at worst misleading”.

## 2.2 Missing Information Principle and EM

We have seen that the idea of ‘filling in’ for missing data is intuitive and has a long history. One of the key papers that contributes to the understanding of what to fill in is Orchard and Woodbury (1972). They indicate that the missing data ‘should be replaced by sample values from an appropriate distribution function’. But which distribution is appropriate? And what sample values?

To facilitate our discussion, we first define some notation. Denote the complete data by \( x \), the observed data by \( y \) and their corresponding sample spaces by \( \mathcal{X} \) and \( \mathcal{Y} \). The complete data \( x \) defines a point in the \( \mathcal{X} \) space but is not directly observed. Assume a many-to-one mapping of the complete data \( x \) to the observed data \( y \). For example, in the case where two groups of integer
data are aggregated into one by addition, an observed value corresponds to
the many plausible values from each of the two individual groups that sum
to that observed value. Hence, after observing $y$, $x$ is only known to lie in
the subset of $X$ defined by a mapping $\pi : x \rightarrow y$.

Suppose the complete data $x = \{x_1, \ldots, x_n\} \in X$, not necessarily independent and identically distributed or even independent, are realizations of random variables from some distribution $P(dx | \theta^*)$ of the family $\{P(dx | \theta) : \theta \in \Theta\}$ of distributions on $X$ which defines the complete model. Assume these distributions have densities $f(x | \theta)$ with respect to some $\sigma$-finite reference measure denoted by $dx$ for brevity. The distribution of $y$ is the image under $\pi$ of the distribution $P(dx | \theta^*)$. Usually, these distributions have densities $g(y | \theta)$ with respect to some reference measure $dy$ on $Y$.

In order for the notation to be more transparent, write $x = (y, z)$ and assume that $X = Y \times Z$, $dx = dy \times dz$ and $\pi(y, z) = y$. The space $Z$ consists of all of the possible unobserved samples $z$.

Denote the density of $y$ by $g(y | \theta)$, then

$$g(y | \theta) = \int_Z f((y, z) | \theta) \, dz.$$  

Further let $k(z | \theta, y)$ denote the conditional distribution of $z$ given $y$, whose density with respect to $dz$ is

$$k(z | \theta, y) = f(x | \theta) / g(y | \theta), \quad (2.4)$$

for $g(y | \theta) \neq 0$. Having set up a notation, we return to the question of which distribution should be used when 'filling in' for the missing data. We paraphrase the original Orchard and Woodbury (1972) paper.

The proper distribution to sample for the missing data then is the conditional distribution $f(z_i | \theta, y_i)$, but $\theta$ is unknown so that some estimated value $\hat{\theta}$, must be used. One could draw many samples from the distribution $f$ and from these completed data samples obtain the distribution of the parameter estimates due to the missing data. Call this distribution MID($\hat{\theta}, Z$). If this distribution is asymptotically normal, then the mean will be the obvious statistic to use to provide an estimate in the presence of missing data. If this mean value should be $\hat{\theta}$, then the estimate
has not been affected by the assumed missing data distribution. That is the missing data tells you nothing. This interpretation of the principle is due to Jacquez. The remaining part of the missing information principle is to equate the mean of the \( \text{MID}(\hat{\theta}, Z) \) to \( \hat{\theta} \), or take some action equivalent to this.

A most important contribution of the Orchard and Woodbury paper is in “taking action equivalent to” equating the mean of \( \text{MID}(\hat{\theta}, Z) \) to \( \hat{\theta} \). Under a classical likelihood-based inference framework, they show that instead of solving the usual likelihood equation, that is, setting the score function of the available observed data to zero, we could have obtained the same solution by setting the expected score function of the unavailable complete data to zero, the expectation being taken with respect to the ‘proper’ distributions \( k(z \mid \theta, y) \) (or \( f(z_i \mid \theta, y_i) \)) in Orchard and Woodbury notation. Since \( \theta \) is not known, we use an iterative algorithm. The solution, under regularity conditions, is the MLE. The authors coin the phrase missing information principle to characterize such maximum likelihood estimation with incomplete data.

Another important contribution of the Orchard and Woodbury paper is that it proposes iterative schemes for solving the expected score function equations of several examples including structural missing data problems. However, no convergence result is provided and the examples they analyze are not extensive.

The missing information principle is further elucidated by Beale and Little (1975). However, the desirable properties of the iterative algorithm are only made explicit and formal in the seminal paper by Dempster, Laird and Rubin (1977), referred to as D-L-R hereafter. The D-L-R paper integrates previous work, broadens the definition of incomplete data to include several new examples with structural missing information, and coins the name EM (Expectation-Maximization) for the algorithm. Despite the fact that it is built upon previous work and contains an error in its proof of convergence, the D-L-R paper is still the most widely cited paper as far as missing data are concerned. The D-L-R paper shows that the missing information principle is equivalent to EM, and is also intimately related to the self-consistency principle introduced by Efron (1967). In fact, the self-consistency principle can be regarded as a specific instance of EM.
2.3 The EM Algorithm

With the notation of the last section, we take the logarithm of both sides of (2.4) and rearrange terms to obtain

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

(2.5)

where $\ell_c(\theta; x) = \log f(x | \theta)$, $\ell_{\text{obs}}(\theta; y) = \log g(y | \theta)$ and $\ell_{\text{cond}}(\theta; x | y) = \log k(z | \theta, y)$. The basic idea of the EM algorithm is to replace maximization of the loglikelihood of the observed data, $\ell_{\text{obs}}(\theta)$, with the successive maximizations of the conditional expectation $Q(\theta | \theta^{(m)})$ of the complete loglikelihood function $\ell_c(\theta)$ given the observations $y$ and the current fit $\theta^{(m)}$ of the parameter. To be specific,

$$Q(\theta | \theta^{(m)}) = E_{\theta^{(m)}}(\ell_c(\theta; x) | y).$$

Given the current approximation $\theta^{(m)}$ to the MLE based on the observed data $y$, the $(m+1)$-th EM iteration is conducted in two steps.

1. E step: Compute $Q(\theta | \theta^{(m)})$.

2. M step: Update $\theta^{(m)}$ by computing the maximizer $\theta^{(m+1)}$ of $Q(\theta | \theta^{(m)})$.

This iterative process is repeated until convergence is apparent. The EM algorithm has the basic property that each iteration increases the loglikelihood function of the observed data, i.e.,

$$\ell_{\text{obs}}(\theta^{(m+1)}) \geq \ell_{\text{obs}}(\theta^{(m)}).$$

Because an estimate of the parameter in the $(m+1)$-th iteration is some function of the parameter estimate of the previous $m$-th iteration, we denote this function or the EM operator, that maps $\theta^{(m)}$ to $\theta^{(m+1)}$ by $T_{EM}$, that is,

$$\theta^{(m+1)} = T_{EM}(\theta^{(m)}).$$

(2.6)

Under rather general conditions, $\ell_{\text{obs}}(\theta^{(m+1)}) = \ell_{\text{obs}}(\theta^{(m)})$ if and only if $\theta^{(m)}$ is a fixed point of the equation

$$T_{EM}(\theta) = \theta.$$  

(2.7)
A detailed study of the convergence properties of the sequence \( \{ \theta^{(m)} \} \) generated by EM with starting position \( \theta^{(0)} \) can be found in Wu (1983) and Redner and Walker (1988). Under sufficient regularity conditions, \( \theta^{(m)} \) converges to a stationary point of \( \ell_{\text{obs}}(\theta) \), i.e., a point \( \theta \in \Theta \) such that the Jacobian matrix \( D\ell_{\text{obs}}(\theta) = 0 \). When \( \ell_{\text{obs}}(\theta) \) has several stationary points (local maxima, local minima, saddle points), the \( \theta^{(m)} \)'s can converge to any of the local maxima or saddle points, depending on the initial position \( \theta^{(0)} \).

Although the iterative algorithm described above is generally known as EM, it would be fair to say that it is exactly the missing information principle. Beale and Little (1975) define the missing information principle as follows: \( \theta \) should be estimated by a solution to the fixed point equation (2.7). Despite technicalities in the difference between fixed points and stationary values, the solutions of the fixed point equations in general satisfy the likelihood equations found by setting the expected score function of the complete data loglikelihood to zero, which is the missing information principle implicitly defined by Orchard and Woodbury (1972).

When the complete data density can be put down in exponential families form, Sundberg (1974) notes that EM takes a strikingly simple form. In such a situation, EM allows us to use the intuitive idea of 'filling in' for the missing data. We can ‘fill in’ for the sufficient statistics using the their expected values given the observed data and a guess of the current parameter. To illustrate this idea, we reanalyze Example 2.1 and show that such ‘filling in’ by EM indeed gives us the correct answer. We discuss the exponential family in depth in Chapter 3.

Example 2.1. (continued) The complete data is \( x_i = (\beta_i, y_i), i = 1, \ldots, n; \) the \( \beta_i \)'s are missing. But now instead of ‘filling in’ for each individual missing data \( \beta_i \) with \( E(\beta_i \mid y, \theta^{(m)}) \), we ‘fill in’ for the sufficient statistics \( \sum_{i=1}^{n} \beta_i^2 \) with \( E(\sum_{i=1}^{n} \beta_i^2 \mid y, \theta^{(m)}) \):

\[
E(\sum_{i=1}^{n} \beta_i^2 \mid y, \theta^{(m)}) = \sum_{i=1}^{n} \left\{ E(\beta_i \mid y, \theta^{(m)}) \right\}^2 + \sum_{i=1}^{n} \text{var}(\beta_i),
\]

which is the E-step. From (2.1) \( \sum_{i=1}^{n} \text{var}(\beta_i) = n \frac{\theta^{(m)} \tau^2}{\theta^{(m)} + \tau^2} \). After filling in for the sufficient statistics the M-step yields

\[
\theta^{(m+1)} = \frac{\theta^{(m)}^2}{n(\theta^{(m)} + \tau^2)^2} \sum_{i=1}^{n} y_i^2 + \frac{\theta^{(m)} \tau^2}{\theta^{(m)} + \tau^2},
\]

18
from which the corresponding fixed point equation is

\[ \theta = T(\theta) = \frac{\theta^2}{(\theta + \tau^2)^2} s^2 + \frac{\theta \tau^2}{\theta + \tau^2}. \]  

(2.8)

Using a geometric technique similar to the one used in analyzing (2.2), we obtain from (2.8):

\[ \hat{\theta}_{EM} = \begin{cases} s^2 - \tau^2, & \text{if } s^2 > \tau^2, \\ 0, & \text{if } s^2 \leq \tau^2, \end{cases} \]

which is exactly the MLE, as promised by EM. ◊

2.4 Limitations of EM

The widespread use of EM since the D-L-R paper is phenomenal. Actually, the D-L-R paper is reported to be one of the six mostly cited paper in the statistics literature according to a recent Statistical Science survey (Stigler 1994). Besides the usual application to missing, censored, truncated and grouped data, EM has been applied in mixture analysis, hyperparameter estimation, variance component models, factor analysis, latent variable analysis, empirical Bayes models, time series estimation, estimation of gene frequencies, survival analysis. Newer applications include image analysis such as positron emission tomography (PET) (Vardi, Shepp and Kaufman 1985), filtering method (Dembo and Zeitouni 1986), hidden Markov models (Baum 1970), adjusting census undercount (Cressie 1992). Recent applications in the neural network literature include object recognition (Utans 1993), hierarchical mixture of experts (Jordan and Jacobs 1994) and statistical physics (Yuille, Stolorz and Utans 1994). Several variants of EM have also been proposed, such as Smoothed EM (Silverman, Jones, Wilson and Nychka 1990), and EM on inverse problems (Vardi and Lee 1994).

One appeal of EM is that at each iteration, the likelihood is guaranteed to increase. Another attractive feature is that it is relatively easy to program. The estimate it derives, namely the maximum likelihood estimate (MLE), also has been investigated extensively and found to have desirable large sample properties.

In spite of its many appealing features, the EM algorithm described in Section 3 has several drawbacks. First, it is not guaranteed to converge. If
it does, it can converge to a local maximum or a saddle point of the loglikelihood function. In many instances, the limiting point depends strongly on its starting position.

Second, the E-step can be computationally difficult. The censored time series discussed in Chapter 1 is such an instance. In other complex models such as hidden Markov models, the E-step or both the E and the M steps can become mathematically intractable. See Qian and Titterington (1991) for a discussion of such problems.

Another instance where the E-step may become infeasible is in estimation in hierarchical models, where the parameters are themselves considered as random samples from a distribution. They are treated as missing data in the EM setup. When the dimension of the random parameter is high, EM requires performing sometimes prohibitively numerous high dimensional integrations. The following example illustrates this fact.

**Example 2.2. NAEP data.** (Thomas 1992; Thomas and Ip 1994)

The National Assessment of Educational Progress (NAEP) uses a multivariate multiple regression model with $p$ latent variables representing the proficiency of examinees in several scales for reporting purposes. For example, in the NAEP mathematics assessment, the proficiency scales include *Numbers and Operations, Measurement, Geometry, and Algebra and Functions*. See Beaton and Zwick (1992) for an overview of NAEP.

Suppose the proficiency scales, denoted by $\theta^T$, is a vector of length $p$ and we have a sample of $n$ examinees. The $i$-th examinee’s scales $\theta_i$ are modeled as a sample from $\mathcal{N}(\Gamma^T x_i, \Sigma)$, where $x_i$ is a vector of $m$ predictors of demographic and educational characteristics, $\Gamma$ and $\Sigma$ are the parameters of interest. The variables $\theta_i$, however, are not observed directly. What we observe are the item responses $y_{ij}$ of each examinee. NAEP uses a hierarchical conditional independent model that assumes that, conditional on $\theta_i$, the responses of the $i$-th examinee are independent with a distribution denoted by $f(y_{ij} \mid \theta_i) = \prod_{j=1}^m f(y_{ij} \mid \theta_i)$.

Under the above model, the likelihood function for $(\Gamma, \Sigma)$ is

$$\prod_{i=1}^n \int f(y_{ij} \mid \theta_i) \phi(\theta_i; \Gamma^T x_i, \Sigma) \, d\theta_i.$$

By treating the unobserved variables $\theta_i$’s as missing data, we can use the EM algorithm to obtain an estimate for $(\Gamma, \Sigma)$. The E-step involves computing
the first two moments, namely \( \bar{\theta}_i = E(\theta_i \mid X, Y, \Gamma^{(m)}, \Sigma^{(m)}) \) and \( \text{var}(\theta_i \mid X, Y, \Gamma^{(m)}, \Sigma^{(m)}) \), and the M-step computes

\[
\Gamma^{(m+1)} = (X^T X)^{-1} X^T (\bar{\theta}_1, \ldots, \bar{\theta}_n)^T,
\]

\[
\Sigma^{(m+1)} = \frac{1}{n} \left\{ \sum_{i=1}^n \text{var}(\theta_i \mid X, Y, \Gamma^{(m)}, \Sigma^{(m)}) + \sum_{i=1}^n [\bar{\theta}_i - (\Gamma^{(m+1)})^T X_i][\bar{\theta}_i - (\Gamma^{(m+1)})^T X_i]^T \right\}.
\]

Each E-step involves the evaluation of a \( p \)-dimensional integral (\( p \) ranges from 1 to 6 depending on the subject content). To illustrate the scope of computation, we suppose \( p = 6 \) and \( n = 10,000 \). The number of iterations required in EM is in the order 100 (in practice, it ranges from 100 to 1,000). The number of parameters to be estimated is 27, including elements in the mean and the covariance matrix. Therefore the number of six-dimensional integrals to be evaluated is \( 6 \times 10,000 \times 100 \times 27 = 162 \) million. As a consequence, we may want to use approximate methods to circumvent this computational problem. ♦

In Chapter 5, we present another example on SAT data where billions of 14-dimensional integrations are required. It would be infeasible to implement EM in these situations.

### 2.5 Filling in Stochastically

As observed by Efron (1977) in his comment on the D-L-R paper, the underlying identity driving EM is the Fisher identity:

\[
\frac{\partial}{\partial \theta} \ell_{\text{obs}}(\theta; y) = E_{\theta} \left( \frac{\partial}{\partial \theta} \ell_{\text{c}}(\theta; x) \mid y \right).
\]  \hspace{1cm} (2.9)

The likelihood equation for the observed data is formed by setting the left hand side of (2.9) to 0. Since the left hand side of (2.9) is difficult to solve, EM solves the right hand side of the equation instead. The complete loglikelihood, \( \ell_{\text{c}}(\theta; x) \) is easier to deal with.

A key ingredient in the identity (2.9) is the conditional distribution \( k(z \mid \theta, y) \) (2.4). To exploit this notion, we take a different perspective. Instead of performing a deterministic substitution for the missing data, we use a stochastic ‘filling in’ scheme. A single sample is drawn from the distribution
(2.4) to 'fill in' for the missing data, after which we proceed as if we have the complete sample.

This idea is called Stochastic Imputation Principle (Broniatowski, Celeux and Diebolt 1983; Celeux and Diebolt 1985, 1987). We now use the terms 'fill in' and 'impute' interchangeably.

Suppose that there is an initial guess \( \theta^{(0)} \) of the parameter \( \theta \), the conditional density (2.4) would be completely known. A sample is drawn from this conditional density to impute for the missing data in order to make the data complete. Then we proceed to seek for an appropriate estimate for the complete data. Denote the updated estimate by \( \theta^{(1)} \). The initial guess \( \theta^{(0)} \) is then replaced by \( \theta^{(1)} \) and we draw again from (2.4). The procedure is iterated until the process stabilizes.

The 'filling in' procedure for missing data from this vantage point is radically different from the traditional viewpoint. The outcome of this procedure is not a point estimate but rather is an array of estimates (for example, MLE's) corresponding to each of the pseudo-complete sample that we artificially created.

We now show how such a procedure works in Example 2.1.

**Example 2.1. (continued).**

Recall that the missing data \( \beta_i \) has a Gaussian conditional density. Therefore, to create a pseudo-complete data set at the m-th iteration, we draw a sample \( \beta_i \) from (2.1),

\[
\beta_i = \frac{\theta^{(m)}}{\sqrt{\theta^{(m)} + \tau^2}} y_i + \varepsilon_i \sqrt{\frac{\theta^{(m)} + \tau^2}{\theta^{(m)} + \tau^2}},
\]

where \( \varepsilon_i \sim \mathcal{N}(0,1) \). Squaring both sides of (2.10), averaging over \( i \), and rearranging terms, we obtain the MLE in the following form,

\[
\theta^{(m+1)} = \frac{\theta^{(m)} + \tau^2}{n(\theta^{(m)} + \tau^2)^2} \sum_{i=1}^{n} y_i^2 + \frac{2c}{n} \sum_{i} y_i \varepsilon_i + \frac{\theta^{(m)} + \tau^2}{n(\theta^{(m)} + \tau^2)} \sum_{i} (\varepsilon_i^2 - 1),
\]

where \( c = \tau \theta^{(m)}/(\theta^{(m)} + \tau^2)^{3/2} \).

The third term in (2.11) is \( O(1/\sqrt{n}) \) and is distributed normally because each \( \varepsilon_i \) is distributed \( \mathcal{N}(0,1) \). The fourth term in (2.11) is also \( O(1/\sqrt{n}) \), since \( \sum (\varepsilon_i^2 - 1)/\sqrt{n} \) has mean 0 and is asymptotically normal. The first two terms constitute the EM operator in (2.7).
When \( n \) is large, (2.11) can be approximated by a normal distribution whose variance decreases in the order of \( 1/n \). We prove in Chapter 4 that the when \( m \) is sufficiently large, the process (2.11) stabilizes and the average of \( \theta^{(m)} \) differs from the MLE by an \( O(1/n) \) term. ♦

From the above example, we see that when imputations are carried out correctly, even in a stochastic manner, the array of pseudo-complete data MLEs would 'cover' the MLE of the observed data in an almost unbiased way.

The above presentation is heuristic, and concepts such as ‘stable’, ‘cover’, etc. need to be defined more rigorously. This is done in Chapter 3.
Chapter 3

Stochastic EM

In Chapter 2 we have seen how missing data can be imputed with stochastic draws from the conditional distribution of the missing data given the observed data and the current model to form a pseudo-complete data. We focus on using the MLE as the estimate for each pseudo-complete data set. The algorithm is called Stochastic EM when the imputation and the maximization are steps performed iteratively. The basic concepts of Stochastic EM are described in Section 1 and in Section 2 we outline results concerning the ergodic properties of the chain generated by the estimates at each Stochastic EM step. Several examples are developed to illustrate these results in Sections 3 and 4.

3.1 Stochastic EM

A statistical model is used to explain data which are partially observed. Based upon knowledge of the current model and the observed data, the stochastic imputation procedure in Chapter 1 provides a pseudo-complete data set, which is easier to analyze.

Various estimation methods can be used to derive point estimates of the parameters using the full stochastically completed data. Nevertheless, we use maximum likelihood estimation for most of our discussion. After the parameters are updated using the MLE of the pseudo-complete data, a new step of stochastic imputation is invoked. The entire procedure is iterated for a sufficient number of times. Under specific conditions, the array of
estimates corresponding to each pseudo-complete data forms a Markov chain that converges to a stationary distribution. We call this algorithm Stochastic EM (Broniatowski, Celeux and Diebolt 1983; Celeux and Diebolt 1985). An example will clarify the exposition of the Stochastic EM algorithm and reveal some of its advantages over EM.

**Example 3.1. Murray data.**

This example is originally reported in Murray (1977) and studied by Tanner and Wong (1987). The data set consists of 12 observations, \((x_i, y_i)^T, i = 1, \ldots, 12:\)

\[
\begin{array}{cccccccccccc}
  x & 1 & 1 & -1 & -1 & 2 & 2 & -2 & -2 & * & * & * \\
  y & 1 & -1 & 1 & -1 & * & * & * & 2 & 2 & -2 & -2 \\
\end{array}
\]

Suppose that the observations are generated from a bivariate normal distribution with a mean zero, a common variance \(\sigma^2\) and a correlation \(\rho\). The parameters of interest are \(\theta = (\sigma^2, \rho)\).

Apparently, using a point estimate naively would be misleading. Tanner and Wong (1987) use a Bayesian methodology called Data Augmentation and obtain a posterior distribution for \(\rho\) that exhibits bimodality. To apply the Stochastic EM to this example, we start with a randomly chosen value of \(\rho^{(0)} = 0.46, \sigma^{(0)} = 5\). Then for each missing \(x_i\), we impute with a draw from the distribution

\[
\mathcal{N}(\rho^{(0)} y_i, (\sigma^{(0)})^2(1 - (\rho^{(0)})^2)).
\]

Missing values of \(y_i\) are imputed in a similar fashion. The MLE \(\theta^{(1)}\), for the pseudo-complete data is computed in the usual way. This process is iterated for 1,000 times.

Figure 3.1 display graphically the values of the 1,000 MLEs corresponding to the respective pseudo-complete data sets.

The Stochastic EM gives estimates that cluster around \(\rho = 0.7\) and \(\rho = -0.7\), whereas estimates of \(\sigma^2\) cluster around 2.8. By averaging the values of \((\sigma^{(m)})^2\) over the last 200 iterations, we obtain an estimate of \(\sigma^2\) as being 2.83. In this example, the likelihood actually has a saddle point at \(\rho = 0\), \(\sigma^2 = 2.5\), and has two maxima of equal loglikelihoods at \(\rho = 0.5\) and \(-0.5\), \(\sigma^2 = 2.667\). Wu (1983) uses this data set as an example to illustrate how EM could converge to a stationary value which is not a local maximum. The slightly smoothed histograms of \(\rho\) and \(\sigma^2\) are plotted in the graphs in Figure 3.2 (a) and Figure 3.2 (b).
Figure 3.1: Stochastic EM on Murray Data. Plot of variance versus correlation.

One thousand iterations appears to be sufficient for the chain to reach stationarity. Various starting values were tried but the results were almost identical.

Example 3.1, in spite of its artificial nature, provides an illustration of how the Stochastic EM ‘spends more time’ in regions that correspond to the more plausible values of the missing data. Stochastic EM recovers a bimodal structure that could not be recovered by a deterministic algorithm such as EM.

This property of Stochastic EM may lead to an improvement over the deterministic EM in terms of convergence to a ‘good’ stationary point. In mixture models, for examples, many researchers try to modify the EM in order to handle multiple maxima which are often encountered in the log-likelihood surface. Besides the issue of multiple maxima, there is also the
problem that EM is not even guaranteed to converge to a local maximum. It converges to a stationary point closest to its starting value and that particular point could be a saddle point (Wu 1983; Redner and Walker 1984). The convergence of EM is essentially governed by the properties of the mapping
\( T_{EM} : \theta^{(m)} \rightarrow \theta^{(m+1)} \) which is non-linear in general.

To study the properties of the Markov chain defined by a general function
\( \theta^{(m+1)} = T(\theta^{(m)}) \), recall some basic facts concerning the simple linear model
\( \theta^{(m+1)} = B \theta^{(m)} \), where \( B \) is a \( 2 \times 2 \) matrix. Clearly 0 is a fixed point. Denote the eigenvalues of \( B \) by \( \lambda_1 \) and \( \lambda_2 \) (\( \lambda_1 \neq \lambda_2 \)), then there can be several different trajectories of the path of \( \theta^{(m)} \) on the \( \mathbb{R}^2 \) plane. When \( \lambda_1, \lambda_2 \) are real, there may exist stable fixed points (when \( |\lambda_1| < 1 \) and \( |\lambda_2| < 1 \)) or saddle points (when \( |\lambda_1| < 1 \) and \( |\lambda_2| > 1 \)). On the other hand, when both \( |\lambda_1| \) and \( |\lambda_2| \) are greater than unity, the sequence \( \theta^{(m)} \) moves away from the origin instead of being attracted toward it.

When \( T \) is non-linear, the behavior of \( \theta^{(m)} \) could be very erratic, and the process can become almost indistinguishable from a stochastic process that exhibits chaos. Fortunately, \( T \) is generally well behaved in most applications in the context of EM. However, even when the modulus of an eigenvalue of \( DT \), the first derivative of \( T \) at a fixed point is less than but close to one, the behavior of \( T \) can still be problematic. The fixed point is referred to as
'unstable' in such a case. On the other hand, a fixed point $F$ is referred to as 'stable' when the modulus of the eigenvalues of $DT$ at $F$ are all significantly less than one. When the EM algorithm is started close to an unstable fixed point, it often takes many iterations to converge.

If a slight perturbation is added to the mapping $T_{EM}(\theta^{(m)})$, as we do in Stochastic EM, then the points transversed by the Stochastic EM chain evade those unstable fixed points in a relatively short time and the iterations 'spend more time' around fixed points that are stable. Whereas EM necessarily terminates in the first stationary point it encounters, Stochastic EM does not terminate in any stationary point. Roughly, it spends more time around stable fixed points with modulus of eigenvalues significantly less than one and spends less time with unstable points with modulus of eigenvalues close to one. For fixed points with modulus of some eigenvalues larger than 1, EM or Stochastic EM may not converge. The above heuristic, of course, requires formal mathematics to support. This should be an important area of further research.

We must remark here that the Stochastic EM in general does not literally 'add' a perturbation to $T_{EM}(\theta^{(m)})$. In fact, the Stochastic EM iterations take the form of a stochastic difference equation

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

(3.1)

We show in Chapter 4 that for the exponential family, (3.1) actually can be decomposed into an additive equation under fairly general conditions:

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

(3.2)

where $T : R^k \rightarrow R^k$ is exactly the EM operator. Note that (3.2) forms the basis of many time series models. For example, a simple linear autoregressive model of order one takes the form

$$X_{t+1} = \rho X_t + \varepsilon_{t+1}.$$

(3.3)

The operator $T$ in (3.2) is sometimes referred to as the skeleton of the dynamic process (3.2) (Tong 1990). Models of the form (3.2) have been studied by a number of authors under the title of non-linear time series (Tweedie and Meyn 1992; Tong 1990; Diebolt and Guegan 1990; Diebolt and Guegan 1993; Doukhan 1994).
3.2 Ergodicity of Stochastic EM

Our first main concern about the chain \( \{\theta^{(m)}\} \) in (3.2) is stationarity. As pointed out by Tong (1990), there is an intimate relation between stationarity and the stability of the operator \( T \). For example, in model (3.3), when \(|\rho| > 1\), the origin is strongly unstable and (3.3) is not strictly stationary. This section describes the pertinent conditions for stationarity of the Markov chain \( \{\theta^{(m)}\} \).

3.2.1 Markov process preliminary

Although stationarity has been studied in a general context, we only consider theories pertinent to our present context, namely the stochastic process generated in Stochastic EM. From (3.1) it is clear that the estimate derived from the \((m+1)\)-th iteration, \( \theta^{(m+1)} \), depends on the value of the \( m \)-th estimate, \( \theta^{(m)} \). Hence the estimated parameter values corresponding to each pseudo-complete data form a chain which is Markov. When the associated EM operator is well behaved, we expect the sequence \( \{\theta^{(m)}\} \) generated by the Stochastic EM to be ergodic and the stationary distribution to be unique.

First we set up some basic notation and definitions. Let \( \mathbb{Z} \) and \( \mathbb{R} \) denote the set of integers and real numbers respectively and let \( T = \mathbb{Z} \) or \( \mathbb{R} \).

**Definition 3.1** A Markov process \( \{\theta^{(m)}\} \), \( m \in T \), is defined by a measure space \( (E, \mathcal{E}) \) and a probability space \( (\Omega, \mathcal{A}, P) \) via a transition kernel \( K^t_s(x, dx') \) such that \( K^t_s(x, F) = P(\theta(t) \in F \mid \theta(s) = x) \) for any \( x \in E \), \( F \in \mathcal{E} \).

**Definition 3.2** A transition probability kernel \( K^t_s(x, dx') \) on the measure space \( (E, \mathcal{E}) \) where \( t, s \) are indexed by \( T \), is defined as a family of probability kernels \( \{K^t_s\}_{s \leq t} \) such that

\[
K^t_s = K^u_s K^t_u
\]

for \( s \leq u \leq t \in T \).

The \( n \)-step transition kernel, when \( T \) is the set of non-negative integers, is denoted by \( K^n(x, F) = P(\theta^{(n)} \in F \mid \theta^{(0)} = x) \). For simplicity, denote
$K^1$ by $K$. Hence, $K(x, dx') = P(\theta^{(1)} \in dx \mid \theta^{(0)} = x)$. Also, the phrases 'transition kernel' and 'Markov kernel' are used interchangeably.

Because $\theta^{(m)} \in \mathbb{R}^k$ in many applications, we deal with continuous state space in general. To generalize the concept of irreducibility from discrete state space to continuous state space, we define $\lambda$-irreducibility (on $\mathbb{R}^k$) in the following way.

**Definition 3.3** A Markov chain $\{\theta^{(m)}\}$ is said to be $\lambda$-irreducible if, for some finite measure $\lambda$ on $(\mathbb{R}^k, \mathcal{B})$, $\mathcal{B}$ being the $\sigma$-algebra of Borel sets in $\mathbb{R}^k$,

$$\sum_{m>0} P(\theta^{(m)} \in A \mid \theta^{(0)} = \theta) > 0$$

for all $\theta$ whenever $\lambda(A) > 0$ and $A \in \mathcal{B}$.

**Definition 3.4** A Markov chain with a transition kernel $K^n(x, .)$, is geometrically ergodic if

$$\|K^n(x, .) - \Psi\|_{TV} \leq A(x)\eta^n, \quad \text{for some} \quad 0 < \eta < 1,$$

for some function $A(x) > 0$ with $\int A(x) \Psi(dx) < \infty$, $\Psi$ a.s.

where $\|\cdot\|_{TV}$ denotes the total variation norm. The distribution $\Psi$ in (3.4) satisfies the invariant equation

$$\Psi(A) = \int K(x, A)\Psi(dx), \quad A \in \mathcal{B}.$$  \hfill (3.5)

If the Markov chain is geometric ergodic, $\Psi$, the limit of $K^n$ is referred to as the stationary distribution of the Markov chain. This stationary distribution is unique by definition.

Geometric ergodicity is more important than ergodicity since the rate of approaching stationarity ought to be fast for the stationarity assumption to be relevant. For details about ergodic properties of Markov chains, see Nummelin (1984).

### 3.2.2 Ergodicity Under Compactness

Ergodicity can be established without major difficulty when the Markov chain $\{\theta^{(m)}\}$ takes values in a compact space. In many cases, it turns out to be
more convenient to first establish the ergodicity of \{z^{(m)}\}, the Stochastic EM chain formed by the ‘filled in’ data at each iteration. The sequence \{z^{(m)}\} is a Markov chain with conditional distribution (density) \(z^{(m+1)} = z'\) given \(z^{(m)} = z\), as its transition kernel (respectively, transition kernel density).

Formally,

\[
K(z, dz') = k\left(dz' \mid y, \hat{\theta}_{ML}(y, z)\right);
\]

respectively,

\[
K(z, z') = k\left(z' \mid y, \hat{\theta}_{ML}(y, z)\right).
\]

If the transition kernels are positive, then the chain \(\{z^{(m)}\}\) is \(\lambda\)-irreducible, aperiodic and it has a unique \(\sigma\)-finite invariant measure up to a multiplicative constant. Since we do not know the existence of an invariant probability distribution, we still have to show that the Markov chain \(\{z^{(m)}\}\) is ergodic. This ergodicity implies useful ergodic properties for the dual Markov chain \(\{\theta^{(m)}\}\).

Suppose that either

- (A1) the \(z^{(m)}\)'s take values in a compact space and its transition kernel has a positive continuous density \(K(z, z')\) with respect to Lebesgue measure,

or

- (A2) the \(z^{(m)}\)'s take values in a finite space and the corresponding transition matrix \([K(z, z')]\) has positive entries,

then the chain \(\{z^{(m)}\}\) is \(\lambda\)-irreducible, aperiodic and geometric ergodic.

Denote the invariant distribution that \(\{z^{(m)}\}\) converges to by \(\pi(dz)\). The essential properties of the chain \(\{z^{(m)}\}\) can be transferred to the sequence \(\{\theta^{(m)}\}\), because of the duality principle of Diebolt and Robert (1993). Actually, we can deduce the geometric convergence of the distributions \(\Psi^{(m)}(d\theta)\) of the random variables \(\theta^{(m)}\) to the stationary distribution \(\Psi(d\theta)\) from the geometric convergence of \(\{z^{(m)}\}\) to the invariant distribution \(\pi(dz)\). Here,

\[
\Psi(d\theta) = \int_{\mathbb{Z}} I\{\hat{\theta}_{ML}(y, z) \in d\theta\} \pi(z) \, dz,
\]

where \(I(.)\) is the indicator function. Alternatively, (3.8) implies

\[
\pi(dz) = \int_{\theta} k(dz \mid y, \theta) \Psi(d\theta).
\]
The above exposition can be summarized by the following proposition.

**Theorem 3.1** Under the assumptions (A1) or (A2):

(i) There exist a constant $C > 0$ and $0 \leq \rho < 1$ such that, for each initial value $\theta^{(0)}$,

$$|| \Psi^{(m)} - \Psi ||_{TV} \leq C \rho^m.$$ 

(ii) For any bounded function $h(\theta)$ and any initial distribution $\Psi^{(0)}$, there exists a constant $C' > 0$ and and $0 \leq \rho \leq 1$ such that

$$|E_{\Psi^{(m)}}[h(\theta)] - E_{\Psi}[h(\theta)]| \leq C' \rho^m.$$ 

**Proof.** We give the proof in the case where the transition kernel has a positive continuous density with respect to Lebesgue measure.

(i) The compact state homogeneous Markov chain $\{z^{(m)}\}$ is $\lambda$-irreducible and aperiodic since $K(z, z') > 0$ for all $z$ and $z'$. Since $K(z, z')$ is also continuous and $Z$ is compact, $\{z^{(m)}\}$ is a Doeblin Markov chain and is therefore uniformly geometrically ergodic and geometrically $\varphi$-mixing. Thus, letting $k(z)$ denote its invariant density, which is a positive and continuous function of $z$, there exist constants $C_1 > 0$ and $0 \leq \rho < 1$ such that

$$||k^{(m)} - k||_1 \leq C_1 \rho^m$$

for any initial density $k^{(0)}(z)$, where $k^{(m)}(z)$ is the density of $z^{(m)}$ and $|| \cdot ||_1$ denotes the $L^1$ norm associated with Lebesgue measure. Now,

$$\Psi(A) = \int_Z I\{\theta_{ML}(y, z) \in A\} k(z) \, dz$$

is a stationary distribution of the Markov chain $\{\theta^{(m)}\}$ and

$$||\Psi^{(m+1)} - \Psi||_{TV} = \sup_A \left| \int_Z I\{\hat{\theta}_{ML}(y, z) \in A\} \left[ k^{(m)}(z) - k(z) \right] \, dz \right|$$

$$\leq \sup_A \int_Z I\{\hat{\theta}_{ML}(y, z) \in A\} \left| k^{(m)}(z) - k(z) \right| \, dz$$

$$\leq ||k^{(m)} - k||_1.$$  

Result (i) then follows.
(ii) Since $h$ is bounded,

$$
|E_{\varphi}^{(m+1)}[h(\theta)] - E_{\varphi}[h(\theta)]| \leq \int_\Theta |h(\theta)| \int_Z I\{\hat{\theta}_{ML}(y,z) \in d\theta\} |k^{(m)}(z) - k(z)| dz
$$

$$
\leq ||h||_\infty \int_Z |k^{(m)}(z) - k(z)| \int_\Theta I\{\hat{\theta}_{ML}(y,z) \in d\theta\} dz
$$

$$
\leq ||h||_\infty ||k^{(m)} - k||_1
$$

$$
\leq ||h||_\infty C_1 \varepsilon^m
$$

(3.12)

according to (i).

\diamond

Note that compactness condition on $\theta^{(m)}$ is not required.

### 3.2.3 Ergodicity under general conditions

In the above section, the duality principle is applied when the domain of $z^{(m)}$ is compact. When neither the domain of $Z$ nor $\theta$ is compact but the transition kernel has a positive continuous density, we can still obtain ergodicity under certain conditions, and the duality principle still applies. Here, we only give conditions on $\theta^{(m)}$ directly. The extension of the following theorems to the ergodicity of $z^{(m)}$ is straightforward and we do not pursue the issue further.

General ergodic results for Markov processes are obtained in Tweedie (1975), Feigin and Tweedie (1985). We state the first theorem here without proof. This theorem is a slight modification of that stated in Chan (1990).

**Theorem 3.2** 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

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

(3.13)

and

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

(3.14)

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

As remarked in Chan (1990) we can interpret $g(\theta)$ as a generalized energy function. The inequality (3.13) asserts that, if the Markov chain starts
outside the compact set $C$, it would, on average, dissipate energy in the next step. The compact set $C$ can be regarded as the center of the state space.

It is known that moments need not exist for stationary dynamic process in general. The following theorem obtained by Tweedie (1983) gives the conditions under which moments exist for the stationary distribution. The statement is a slight modification of that stated in Tong (1990).

**Theorem 3.3** 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 $A$,
- $\sup E[f(\theta^{(m+1)}) \mid \theta^{(m)} = \theta] < \infty, \quad \theta \in A$,
- there exists $0 < \rho < 1$ such that

$$E[f(\theta^{(m+1)}) \mid \theta^{(m)} = \theta] \leq \rho f(\theta), \quad \theta \notin A; \quad (3.15)$$

then

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

Moreover, for some $\omega < 1$,

$$\int_{\mathbb{R}^k} \Psi(d\theta) \sup_{|h| \leq f} \left| E[h(\theta^{(m)}) \mid \theta^{(0)} = \theta] - \int_{\mathbb{R}^k} \Psi(dy) h(y) \right| = O(\omega^m), \quad \text{as} \quad m \to \infty.$$

Inequality (3.15) asserts that if we can find a compact set such that outside it, the jump of the function of $\theta$ from one step to another is not too wild, then the function can be integrated.

Both Theorems (3.2) and (3.3) are used in Chapter 4 to show ergodicity and existence of moments of the Markov chain generated by Stochastic EM within the exponential families framework.

As pointed out by Diebolt and Celeux (1993), the theoretical study of Stochastic EM is difficult in a general incomplete data setting. It relies on existence and consistency results of maximum likelihood methods and convergence results for EM. Apparently, the introduction of the stochastic step yields additional complexity. Two problems are identified:
1. The ergodicity of the Markov chain generated by Stochastic EM, which ensures the existence of a unique stationary distribution. It most cases, this can be established with technical assumptions.

2. The behavior of the stationary distribution, and in particular, its asymptotic behavior and properties as the sample size goes to infinity.

We have outlined results concerning the first issue. The latter issue is more problematic. Exact stationary distributions $\Psi(d\theta)$ are often not trivial to obtain. However, if $\{\theta^{(m)}\}$ is an ergodic Markov chain, an implicit solution is always available and is given by (3.5). Exact stationary distributions can only be obtained in particular cases. Before launching into the general theory, we now give several examples in which the stationary distributions of $\theta^{(m)}$ generated by Stochastic EM can actually be obtained explicitly.

### 3.3 Examples of Stationary Distribution

**Example 3.2. 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 iteration to complete the sample, so that we have an update of $\theta$ given by

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

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.

We now have a simple Markov process. It is a dynamic system

$$
\theta^{(m+1)} = T(\theta^{(m)}) + \frac{\sqrt{k}\sigma}{n}\varepsilon^{(m+1)},
$$

where $T$ is linear and the noise $\varepsilon^{(m+1)}$ are independent and identically distributed as $\mathcal{N}(0, 1)$.  

35
In fact, this 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 Markov chain generated by this kernel converges to the fixed point, or eigenfunction of the equation
\[
\Psi(\cdot \mid y) = \int K(\cdot \mid \tau, y) \Psi(\tau \mid y) d\tau.
\]

Because the kernel \( K(\cdot \mid \tau, y) \) is Gaussian, the solution \( \Psi(\cdot \mid y) \) is also Gaussian, with mean \( \mu \) and variance \( \gamma^2 \) given by the following fixed point equations:
\[
\mu = a\mu + b\bar{y}, \quad (3.16)
\]
\[
\gamma^2 = \frac{k}{n^2} \sigma^2 + a^2 \gamma^2. \quad (3.17)
\]

As a result, the stationary distribution of the Markov chain of estimates generated by Stochastic EM is exactly \( \mathcal{N}(\bar{y}, k\sigma^2/[r(n + k)]) \). Note that the mean is the MLE and the variance depends on the ratio of the number of missing data to the number of observed data, \( k/r \). \( \diamond \)

**Example 3.3. 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 \( Z_i \sim \mathcal{N}(\theta, \sigma_1^2) \), 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
\[
\hat{\theta}_{ML} = \bar{y}.
\]

Treat the signal \( Z_i \) as missing values and apply the Stochastic EM algorithm. Similar to Example 3.2, the Markov kernel of the dynamic system is Gaussian. Hence, using an analogous argument, it is straightforward to show that the stationary distribution of \( \theta \) is exactly
\[
\mathcal{N}\left( \bar{y}, \frac{\sigma_2^2(\sigma_1^2 + \sigma_2^2)}{n(\sigma_1^2 + 2\sigma_2^2)} \right).
\]
Example 3.4. Contaminated bivariate signal.

We extend the above example to the bivariate case. Suppose that for \( i = 1, \ldots, n \), \( Z_i \sim \mathcal{N}(\theta, \Sigma_i) \), 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

\[
\mathcal{N}(A\theta^{(m)} + BY, \frac{1}{n}A\Sigma_1),
\]

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 similar to (3.16) and (3.17).

For the mean \( \mu \), the fixed point equation is

\[
\mu = A\mu + B\bar{y}.
\]  

(3.18)

The solution to (3.18) is \( \mu = \bar{y} \).

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

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

(3.19)

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

\[
\frac{1}{n}A\Sigma_1 = \frac{1}{n}(\Sigma_1 - \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}\Sigma_1)
\]

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

Now write

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

in (3.19) 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

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

Example 3.5. Linear Regression.
Consider the following classical regression problem,
\[
y_i = \sum_{j=1}^{p} k_{ij} \theta_j + \sigma \varepsilon_i, \tag{3.20}
\]
where \( i = 1, \ldots, n \). The random variables \( \varepsilon_i \) are independent and identically distributed as standard normal and \( k_{ij}, \sigma \) are assumed to be known. Denote the \( n \times p \) matrix \( (k_{ij}) \) by \( K \).

The MLE of \( \theta = (\theta_1, \ldots, \theta_p)^T \) has a closed form, namely \((K^T K)^{-1} K^T y\). Therefore, it is not necessary to use EM or Stochastic EM to solve the problem. However, in order to illustrate how EM and Stochastic EM can still be used in such a setting, we recast the problem into a missing data one. The purpose of the exercise is to show that the Markov chain generated from Stochastic EM does have desirable asymptotic properties in this classical problem. In addition, it shows how Stochastic EM provides extra information about the variability of the estimates due to missing data.

The model (3.20) can be reformulated into a missing data problem if we treat \( z_{ij} = k_{ij} \theta_j + (1/\sqrt{p}) \varepsilon_{ij} \) as missing data, where \( \varepsilon_{ij} \sim N(0, 1) \) independently. The observed datum \( y_i \) is the aggregated term \( \sum_{j=1}^{p} z_{ij} \).

This formulation of the problem permits the imputation step in Stochastic EM reconstruct the missing \( z_{ij} \)'s. Alternatively, had we known each individual \( z_{ij} \), \( \theta_j \) can be easily estimated by maximum likelihood:
\[
\theta_j^{(m+1)} = \frac{1}{\sum_i k_{ij}^2} \sum_i k_{ij} z_{ij}^{(m)}, \tag{3.21}
\]
where \( z_{ij} \) were imputed with values \( z_{ij}^{(m)} \) at the m-th iteration. The conditional distribution of each missing value \( z_{ij} \) given \( y \) and \( \theta^{(m)} \) is
\[ \mathcal{N}(k_{ij}\tilde{\theta}_j^{(m)} + \frac{1}{p}(y_i - \sum_k k_{ik}\tilde{\theta}_i^{(m)}), \frac{p-1}{p}\sigma^2). \] (3.22)

Combining (3.21) and (3.22), we obtain the transition kernel, which is defined by the vector equation

\[ \theta^{(m+1)} = \theta^{(m)} + \frac{1}{p}\Lambda K^T(y - K\theta^{(m)}) + \varepsilon^{(m+1)}, \] (3.23)

where \( \Lambda \) is a \( p \)-dimensional diagonal matrix whose diagonal elements are of the form \( \lambda_j = 1/\sum_i k_{ij}^2, \; j = 1, \ldots, p \), and \( \varepsilon^{(m+1)} \) are independent and identically distributed with a common multivariate normal distribution \( \mathcal{N}(0, \frac{p-1}{p}\sigma^2\Lambda) \).

Rewrite (3.23) as

\[ \theta^{(m+1)} = P\theta^{(m)} + b + \varepsilon^{(m+1)}, \]

where \( P = I - (1/p)\Lambda K^T K \) and \( b = (1/p)\Lambda K^T y \). Therefore,

\[ \theta^{(m)} = P^m\theta^{(0)} + P^{m-1}b + \ldots + Pb + b + P^{m-1}\varepsilon^{(1)} + \ldots + P\varepsilon^{(m-1)} + \varepsilon^{(m)}. \]

Note that the modulus of every eigenvalue of the \( p \times p \) symmetric matrix \( P \) are strictly less than one. As \( m \to \infty \), \( P^m\theta^{(0)} \to 0 \), and \( P^{m-1}b + \ldots + Pb + b \to (I - P)^{-1}b \). Furthermore, as \( m \to \infty \), \( (P^{m-1}\varepsilon^{(1)} + \ldots + P\varepsilon^{(m-1)} + \varepsilon^{(m)}) \) tends to the distribution of \( (I - P)^{-1}\varepsilon \), \( \varepsilon \sim \mathcal{N}(0, \frac{p-1}{p}\sigma^2\Lambda) \). Hence

\[ \lim_{m \to \infty} \theta^{(m)} = (I - P)^{-1}b + (I - P)^{-1}\varepsilon. \]

Since \( (I - P)^{-1} = p(K^TK)^{-1}\Lambda^{-1} \), it is straightforward to see that the stationary distribution is a \( p \)-dimensional multivariate normal distribution

\[ \mathcal{N}((K^TK)^{-1}K^Ty, p(p-1)\sigma^2(K^TK)^{-1}\Lambda^{-1}(K^TK)^{-1}). \]

The mean of the stationary distribution is exactly the MLE.

It is well known that when the matrix \( K \) is nearly rank deficient, the least squares estimate or the MLE is unstable. To further our investigation on the effect of the matrix \( K \) on the stationary distribution generated from Stochastic EM, we decompose the matrix \( K \) by the single value decomposition into \( V^TSU \), (e.g. Marshall and Olkin 1979; Golub and van Loan 1987). The matrix \( V \) is of dimension \( n \times n \), \( S \) \( n \times p \) \((p < n)\) and \( U \) \( p \times p \); \( V \)
and $U$ and both orthogonal and $S$ has the singular values of $K$ as its diagonal elements. The variance of the stationary distribution now has the form $\frac{p(p-1)}{\lambda} \sigma^2 U^T (S^T S)^{-1} U \Lambda^{-1} U^T (S^T S)^{-1} U$.

To relate the variance to the singular values, take $\Lambda = \lambda I$, where $\lambda > 0$ is a scalar. Then $U \Lambda^{-1} U^T = (1/\lambda) I$, so that we obtain

$$\frac{p(p-1)}{\lambda} \sigma^2 U^T (S^T S)^{-2} U.$$  

Therefore, when some of the singular values of the matrix $K$ are small, the variance along those directions in the eigen space of $K^T K$ would be large. We illustrate this fact by two simple experiments using simulated data generated under the model (3.20).

In both experiments, $p = 3$ and $n = 30$. In the first experiment, the smallest eigenvalue of the matrix $K^T K$ is about 0.001. The population value of $\theta$ is given by $(\theta_1, \theta_2, \theta_3) = (1.0, 1.5, 2.0)$; and $\sigma^2 = 0.05$. In the second experiment, the same population values of the parameters are used but the eigenvalues of $K^T K$ are all quite far away from 0.

We use the term ‘plausible region’ to describe the region visited by the Markov chain $\{\theta^{(m)}\}$ since each $\theta^{(m)}$ corresponds to a plausible complete data set $(y, z^{(m)})$. Figure 3.3 (a), (b) shows the plausible regions in the $(\theta_2, \theta_3)$
Figure 3.4: EM iterations when X is nearly singular (a): distance between successive iterations; (b): drifting of EM iterations.

space visited by the Stochastic EM Markov chains in the Experiments 1 and 2 respectively. Both chains were started from some random values. A diffuse plausible region with several clusters can be seen for a near singular matrix $K$ in the first experiment. For the second experiment, the plausible region forms a tightly clustered cloud.

To illustrate the unstable nature of the EM algorithm in Experiment 1, we ran EM for 1,000 iterations. Figure 3.4(a), (b) shows that even the Euclidean distances between estimates of successive iterations are small, one of the regression coefficients, $\theta_3$ actually drifts along without ever converging.

The above example shows how the Markov chain produced by Stochastic EM indeed provides valuable information about the parameter estimates in missing data problems. A graphical display may help us diagnose the extent to which information is lost due to missing data. For example, a high variance observed in the graphical display is indicative of a big loss of information.

In all of the above examples, notice that $\theta^{(m+1)}$ is related to $\theta^{(m)}$ in the form

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

where $\varepsilon^{(m+1)} \sim \mathcal{N}(0, I)$, and $A, b, S$ are all independent of $\theta^{(m)}$. 

41
In the following example, \( S \) is a function of \( \theta^{(m)} \) and the Gaussian kernel method is no longer useful in obtaining stationary distribution. Fortunately, by making asymptotic approximations, we can still derive an explicit form for its stationary distribution.

**Example 3.6. Censored exponential data.**

Consider a simple censored exponential data example. The complete data \( x \) is a random sample of size \( n \) from an exponential distribution with mean \( \theta^* \). The observed data are those smaller than a certain censoring point \( c \). Denote the observed data by \((y_1, \ldots, y_q)\) and the number of censored data by \( r \). Hence \( n = q + r \) and we assume that \( \rho_n = r/n \to \rho^* \neq 0 \) as \( n \to \infty \). The MLE \( \hat{\theta}_n \) in this example can actually be calculated directly:

\[
\hat{\theta}_n = \frac{\rho_n c + (1 - \rho_n)\bar{y}}{1 - \rho_n}.
\]

Consider the Stochastic EM algorithm for this incomplete data problem. By the memoryless property of the exponential distribution, the stochastic imputation step consists of simulating \( z_1^{(m)}, \ldots, z_r^{(m)} \) independently from the exponential distribution with mean \( \theta^{(m)} \). This yields the Stochastic EM updating:

\[
\theta^{(m+1)} = \rho_n \theta^{(m)} + (1 - \rho_n)\bar{y} + \rho_n c + \frac{\rho_n}{n} \sum_{j=1}^r (z_j^{(m)} - \theta^{(m)})
\]

\[
= T(\theta^{(m)}) + \frac{1}{n} \sum_{j=1}^r (z_j^{(m)} - \theta^{(m)}),
\]

(3.24)

where \( T(\theta) = \rho_n \theta + (1 - \rho_n)\bar{y} + \rho_n c \) is the EM operator. Equation (3.24) can be rewritten as

\[
\theta^{(m+1)} - \hat{\theta}_n = \rho_n (\theta^{(m)} - \hat{\theta}_n) + \frac{S_n(\theta^{(m)})}{\sqrt{n}} \eta^{(m+1)},
\]

(3.25)

where \( S_n(\theta) = \rho_n^{1/2} \theta \). The terms \( \eta^{(m+1)} = (1/\sqrt{r}) \sum_{j=1}^r (e_j^{(m)} - 1) \) and \( e_j^{(m)} \), \( j = 1, \ldots, r \) are independent and identically distributed exponential random variables with mean 1.

If \( c \) is fixed then \( S_n^2(\theta) \to \rho^* \theta^2 \) when \( n \to \infty \). The distribution of each \( \eta^{(m)} \) also converges to \( \mathcal{N}(0, 1) \). Thus, the sequence generated by Stochastic
EM is approximately a Gaussian autoregressive process of order one with variance of order $1/n$:

$$
\theta^{(m+1)} \approx T(\theta^{(m)}) + \frac{\sqrt{\rho_n} \theta^{(m)}}{\sqrt{n}} \varepsilon^{(m+1)},
$$

(3.26)

where the $\varepsilon^{(m+1)}$'s, $m \geq 0$, are a sequence of independent and identically distributed $\mathcal{N}(0,1)$ random variables.

The following results concerns with the Stochastic EM Markov chain generated by (3.25) if $n$ is sufficiently large.

Theorem 3.4 (i) The Stochastic EM Markov chain is geometrically ergodic. (ii) The mean of its stationary distribution $\Psi_n$ is the MLE, $\hat{\theta}_n$. (iii) The variance of $\Psi_n$ is approximately $\rho_n \hat{\theta}_n^2 / n(1 - \rho_n^2)$. (iv) The stationary distribution $\Psi_n$ is asymptotically Gaussian with mean $\theta^*$ and variance $\rho^* \theta^* /

Although this example is straightforward in its structure, it is instructive to see an alternative proof to the kernel method in obtaining explicit expressions for the mean, the variance and the asymptotic normality of the stationary distribution.

Sketch of the proof of the Theorem:

Assertion (i) is a direct consequence of Theorem (4.3) in Chapter 4.

For (ii): Let $U^{(m)} = \theta^{(m)} - \hat{\theta}_n$. As $n$ is sufficient large, equation (3.26) can be used as an approximation to (3.25). We first take expectation on (3.26) conditional on $\theta^{(m)} = \theta$ and then uncondition by taking expectation with respect to $\theta$ given the observed data $y$. Denote by $E(U \mid y)$ the conditional expectation of the random variable $U$ given the observations, we obtain

$$
E\left(\left|U^{(m+1)}\right| \mid y\right) = \rho_n E\left(\left|U^{(m)}\right| \mid y\right).
$$

(3.27)

Since $0 \leq \rho_n < 1$, it follows that

$$
\lim_{m \to \infty} E\left(\left|U^{(m)}\right| \mid y\right) = 0.
$$

(3.28)

Rewrite (3.26) as

$$
U^{(m+1)} = \rho_n U^{(m)} + \left(\frac{\sqrt{\rho_n} U^{(m)}}{\sqrt{n}} + \frac{\sqrt{\rho_n} \hat{\theta}_n}{\sqrt{n}}\right) \varepsilon^{(m+1)}.
$$

(3.29)
Squaring both sides of (3.29), taking expectation conditional on \( \theta^{(m)} = \theta \), and then unconditioning again as in (3.27), we obtain

\[
E \left( \left| U^{(m+1)} \right|^2 \mid y \right) = \rho_n^2 E \left( \left| U^{(m)} \right|^2 \mid y \right) + \frac{\rho_n E \left( \left| U_{n}^{(m)} \right|^2 \mid y \right)}{n} + \frac{\rho_n \hat{\theta}_n^2}{n}. \tag{3.30}
\]

The form of (3.30) is \( X^{(m+1)} = \gamma X^{(m)} + O(1/n) \), \( \gamma = (\rho_n^2 + \rho_n/n) < 1 \) for large enough \( n \). This implies that \( \lim_{n \to \infty} E \left( \left| U^{(m)} \right|^2 \mid y \right) = O (1/n) \) and therefore is bounded for large \( n \). Thus, the sequence \( E \left( \left| U^{(m)} \right|^2 \mid y \right) \) is uniformly integrable and this permits the interchange of expectation and \( \lim_{n \to \infty} \) in (3.28) (Williams 1991, Theorem 13.3 and Theorem 13.7):

\[
\lim_{m \to \infty} E \left( \theta^{(m)} \mid y \right) = \text{Mean} (\Psi_n) = \hat{\theta}_n,
\]

from which (ii) follows. In fact, from (3.30), we observe that

\[
\lim_{m \to \infty} E \left( \left| U^{(m)} \right|^2 \mid y \right) = \frac{\rho_n \hat{\theta}_n^2}{n(1 - \rho_n^2 - \frac{\rho_n}{n})}. \tag{3.31}
\]

Start from (3.29) again, we take the fourth power on both sides and use a similar argument as in arriving at (3.30) to show that the fourth moment \( E(\left| U^{(m)} \right|^4 \mid y) \) is bounded, allowing us to interchange \( \lim_{m \to \infty} \) and expectation in (3.31), implying (iii). Finally, (iv) can be proved as in Diebolt and Celeux (1993), since we have the equation:

\[
\sqrt{n} U^{(m+1)} = \rho_n \sqrt{n} U^{(m)} + \sqrt{n} \hat{\theta}_n \eta^{(m)}, \tag{3.32}
\]

where the \( \eta^{(m)} \)'s are approximately independent and identically distributed \( \mathcal{N}(0, 1) \). See also the proof of Theorem 4.7 in Chapter 4.

\[\Diamond\]

### 3.4 Grouped Multinomial Example

So far we have dealt with dynamic systems

\[
\theta^{(m+1)} = T(\theta^{(m)}) + V(\theta^{(m)}) \varepsilon^{(m+1)}
\]
whose skeletons $T$ are linear. The following example shows a case in which the
skeleton is not linear and the stochastic part is also dependent upon $\theta^m$. In
this case, the stationary distribution cannot be developed analytically unless
approximation is made.

**Example 3.7. Grouped multinomial data.**

There are $n$ litters each of 2 animals with 0, 1 or 2 males offsprings. The
categories of 0 and 1 males are grouped together. Thus, the observed data
are $Y_1$, the number of litters with either zero or one male, and $Y_2$, the number
of litters which are all males. The total $Y_1 + Y_2 = n$.

Our objective is to estimate $\theta$, the proportion of males in the population.
We start with the assumption that the number of males for each litter follows
a $\text{Binomial}(2, \theta)$ distribution. Let $Z_0$, the number of litters with 0 males,
be regarded as missing data. The complete data are $(Z_0, Y_1, Y_2)$. Had we
observed the complete data, the MLE for $\theta$ would be

$$\hat{\theta} = \frac{y_1 - z_0 + 2y_2}{2n}. \quad (3.33)$$

However, $Z_0$ is not observed. The conditional distribution $Z_0 \mid y_1, y_2, \theta^m$
is $\text{Binomial}(y_1, p^m)$, where $p^m = (1 - \theta^m)^2/(1 - \theta^m)^2$. Because $(Z_0, Y_1, Y_2)$
is the complete sufficient statistic, $Z_0$ is substituted with its conditional
expectation in the E-step in EM. Therefore,

$$\theta^{(m+1)} = T(\theta^m) = \frac{y_1 + 2y_2 - y_1(1 - \theta^m)/(1 + \theta^m)}{2n}. \quad (3.34)$$

where $T$ is the EM operator. The rate of convergence of EM is governed by
$| DT(\hat{\theta}) |$, the Jacobian matrix of $T$ (D-L-R 1977; Louis 1982) at $\hat{\theta}$.
Here, $| DT(\theta) |$ is straightforward to evaluate and it is $y_1/n(1 + \theta^2) < 1$.
Thus EM converges to the fixed point of (3.34) by the fixed point theorem (Ortega,
1990), and the solution is the MLE, $\hat{\theta}_n = \sqrt{y_2/n}$, as promised by EM.

To study the Markov process underlying Stochastic EM, we observe that
the draw $z_0^{(m)}$ at the m-th iteration is from the distribution $\text{Binomial}(y_1, p^m)$.
Substituting into (3.33) gives $\theta^{(m+1)} = (y_1 + 2y_2 - z_0^{(m)})/2n$. When $n$ is sufficiently large, the distribution of $z_0^{(m)}$ is approximately $\mathcal{N}(y_1p^{(m)}, y_1p^{(m)}(1 - p^{(m)}))$. Hence asymptotically, $\theta^{(m+1)}$ is distributed as

$$\mathcal{N}\left(\frac{y_1 + 2y_2 - y_1p^{(m)}}{2n}, \frac{y_1p^{(m)}(1 - p^{(m)})}{4n^2}\right), \quad (3.35)$$

45
where \( p^{(m)} = (1 - \theta^{(m)})^2 / (1 - \theta^{(m)})^2 \). In other words,

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

(3.36)

where \( a = -y_1/2n, b = (y_1 + 2y_2)/2n, S(\theta) = \sqrt{(y_1(1 - \theta)\theta)/(4n(1 + \theta)^2)}. \)

Note that \( y_1/n \to \theta(2 - \theta) \) asymptotically. It is evident from (3.36) that the skeleton of the Markov process, which is the EM operator, is non-linear and the variance of the stochastic noise is dependent upon \( \theta^{(m)} \) as well. This non-linear autoregressive process of order one can not be solved analytically. However, for this example, we show in Chapter 4 that ergodicity and finite moment results are still available. The mean of the stationary distribution differs from the MLE by \( O(1/n) \).

\[\diamondsuit\]
Chapter 4

Stationarity in Exponential Families

In Chapter 3 we discussed the ergodic properties of the Markov chain generated by Stochastic EM. Properties of the Stochastic EM iterations have only been studied in a few limited cases such as finite mixture model with two components (Celeux and Diebolt 1987). In this chapter we further develop new, general results concerning the stationary distribution of the Markov chain, with an emphasis on exponential family. After stating some preliminaries on exponential families and information matrices in Section 1, we show in Section 2 that for exponential families, the mean of the stationary distribution differs from the MLE by $O(1/n)$ under appropriate assumptions. In Section 3, we provide a different approach to analyze the Stochastic EM algorithm for exponential families. We show that the Stochastic EM iterations asymptotically form an additive stochastic difference equation. Conditions necessary for geometric ergodicity under such situations are provided in Section 4. In Section 5, the asymptotic stationary distribution of Stochastic EM is explicitly obtained for the prototypical linear case. Finally, we give results for the general case in Section 6.
4.1 Preliminary

4.1.1 Exponential families

Exponential families are characterized by having probability functions of the form

\[ f(x \mid \eta) = \exp[<\theta(\eta), t(x)> - B(\theta(\eta))]h(x), \]  

(4.1)

or in the canonical form

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

(4.2)

where \( \eta \) is a parameter. Both \( x \) and \( \eta \) are multidimensional in general, \( \theta(\eta) \) and \( t(x) \) are vectors of common dimension \( k \), and \( \langle \cdot, \cdot \rangle \) denotes inner product. In (4.2), \( \theta, t(x) \) are called canonical (or natural) parameters and canonical (or natural) statistic respectively. If we let \( \mathcal{P} \) to be the exponential family of distributions with probability functions (4.1), the smallest \( k \) for which an exponential representation of \( \mathcal{P} \) with \( \theta \) and \( t \) of dimensions \( k \) is possible is the order of \( \mathcal{P} \), and such a representation is said to be minimal. The canonical statistic \( t \) is a sufficient statistic for \( \mathcal{P} \) and it is minimal sufficient if the representation (4.1) is minimal (Lehmann 1988). We shall only deal with minimal sufficient statistics \( t \) in the following discussion.

The probability functions \( f(x \mid \eta) \) are all densities with respect to the same measure \( \mu \), which is typically either the counting measure or Lebesgue measure. Let \( \mathcal{T} \) be the domain of \( \eta \) and \( \Theta \) be the corresponding domain of \( \theta \) for \( \mathcal{P} \), assuming \( \theta = \theta(\eta) \) is a one-to-one mapping. Also let \( \hat{\Theta} = \{\theta : \int h(x) \exp(<\theta, t(x)>)d\mu < \infty\} \), which is a convex subset of \( \mathbb{R}^k \). Then \( \mathcal{P} \) is said to be regular if \( \Theta = \hat{\Theta} \) and if \( \Theta \) is an open subset of \( \mathbb{R}^k \).

Exponential families share a large number of important and useful properties which often make an “incisive statistical analysis feasible” (Barndorff-Nielsen 1982). Here we only outline a few properties which are relevant to our subsequent discussion. For further references on exponential families, Lehmann (1988) and Johansen (1979) explain the basics; Brown (1986), Barndorff-Nielsen (1988) and the references in it provide results on the structure and properties. Geometric interpretations can be found in Efron (1975, 1978).

Three important operations on exponential families lead again to exponential families. First, the distribution of the canonical statistic \( t \) is exponential. (Theorem 4.3, Lehmann (1988 page 35)). If \( X \) and \( Y \) are independently
distributed according to (4.1) with different \( t, A(\cdot), h(\cdot) \), the joint distribution of \( (X, Y) \) is again an exponential family. In particular, if \( X_1, \ldots, X_n \) is a random sample from (4.1), the exponential structure is preserved and the joint density is given by

\[
\exp\left[ \sum_{i=1}^{n} < \theta_i(\eta_i), t_i > -nB(\theta(\eta_i)) \right] h(x_1) \ldots h(x_n).
\]

The first two moments of exponential families under the parameterization (4.2) are determined by the derivatives of the function \( A(\theta) \). For each \( j = 1, \ldots, k \),

\[
E_\theta(t_j) = \frac{\partial}{\partial \theta_j} A(\theta), \quad (4.3)
\]

\[
\text{cov}(t_j, t_k) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} A(\theta).
\]

The conditional inference property is another important property of the exponential family. If \( \theta = (\theta^{(1)}, \theta^{(2)}) \), and \( t = (t^{(1)}, t^{(2)}) \) are similar partitions of \( \theta \) and \( t \), then the conditional distribution of \( t^{(2)} \) given \( t^{(1)} \) is exponential family with \( \theta^{(2)} \) and \( t^{(2)} \) as canonical parameter and statistic, respectively.

The maximum likelihood inference on exponential families is also tractable when the parameter of interest is the canonical parameter. Taking the logarithm on both sides of (4.1) and forming the score function leads to the likelihood equation of the form

\[
\frac{\partial \ell}{\partial \theta} = t - A'(\theta) = 0, \quad (4.4)
\]

where \( \ell = \log f(x \mid \theta) \). By virtue of (4.3),(4.4) amounts to, in short hand form,

\[
E_\theta(t) = t. \quad (4.5)
\]

Provided the domain of the parameter \( \theta \) is convex, if a solution to (4.5) exists, then it is unique and is the MLE \( \hat{\theta} \) (Barndorff-Nielson 1982).

In general, the solution to (4.5) does not have a closed form. Newton-Raphson or quasi-Newton methods such as the Davidon-Fletcher-Powell algorithm are efficient ways to compute \( \hat{\theta} \). The asymptotic optimality properties of the solution to (4.5) can be established under fairly general conditions.
See Sundberg (1974), Brown (1986) and Barndorff-Nielsen (1978) for further discussions on MLE in exponential families. As a remark, note that for an arbitrary parameter $\eta = \phi(\theta)$, the MLE $\hat{\eta} = \phi(\hat{\theta})$ by definition.

In the missing data context outlined in Chapter 1 and 2, where the likelihood of the joint distribution of the complete data $X = (Y, Z)$ is given by (4.2), exponential family inference takes a strikingly simple form.

Analogous to (4.5), the likelihood equation for a regular exponential family in the missing data case is

$$E_\theta(t) = E_\theta(t \mid y). \tag{4.6}$$

Equation (4.6) invites an iterative algorithm to solve for $\hat{\theta}$. Let $H(\theta \mid \theta^{(m)}) = E_\theta(t) - E_{\theta^{(m)}}(t \mid y)$. At the m-th iteration, solve $H(\theta \mid \theta^{(m)}) = 0$ for $\theta$ and update $\theta^{(m)}$ to $\theta^{(m+1)}$. The iterations $\theta^{(m)}$ converges to $\hat{\theta}$ under regular conditions. This iterative algorithm is exactly the EM algorithm.

The iterative steps $\theta^{(0)} \to \theta^{(1)} \to \ldots \to \theta^{(m)}$ can be summarized by the EM operator equation $\theta^{(m+1)} = T_{EM}(\theta^{(m)})$. It turns out that the first derivative, or Jacobian of $T$ is related to the Fisher information matrix of the various loglikelihoods. Now we outline some facts about the information matrix in preparation for obtaining results in Sections 3 and 5.

### 4.1.2 Information

The following theory is generally true for any family of distributions. To begin, rewrite the loglikelihood equation (2.5) as

$$l_c(\theta; y, z) = l_{\text{cond}}(\theta; y, z) + l_{\text{obs}}(\theta; y). \tag{4.7}$$

Take the second derivative in (4.7) with respect to $\theta$:

$$\frac{\partial^2}{\partial \theta^2} l_c(\theta; y, z) = \frac{\partial^2}{\partial \theta^2} l_{\text{cond}}(\theta; y, z) + \frac{\partial^2}{\partial \theta^2} l_{\text{obs}}(\theta; y). \tag{4.8}$$

Multiplying (4.8) by $-1$ and further take expectation with respect to the conditional distribution $k(z \mid y, \theta)$ leads to

$$E \left[ -\frac{\partial^2}{\partial \theta^2} l_c(\theta; y, z) \mid y, \theta \right] = E \left[ -\frac{\partial^2}{\partial \theta^2} l_{\text{cond}}(\theta; y, z) \mid y, \theta \right] + \frac{\partial^2}{\partial \theta^2} l_{\text{obs}}(\theta; y). \tag{4.9}$$
We can interpret equation (4.9) as 'complete information' equals 'conditional information' plus 'observed information'. To simplify notation, write (4.9) as

\[ J_c = J_{\text{cond}} + J_{\text{obs}}. \]  

(4.10)

Note that only \( J_{\text{cond}} \) is the information matrix in the Fisherian sense, that is, the expectation of the negative of the second derivative of the loglikelihood function. The second term on the right hand side of (4.10), \( J_{\text{obs}} \) is the observed information matrix of the observed data. Both \( J_c \) and \( J_{\text{obs}} \) are not information matrices in the true Fisherian sense. However, when we further take expectation with respect to the distribution of \( Y \) on (4.10), we obtain, following the notation in Orchard and Woodbury (1972),

\[ I_X = I_{X|Y} + I_Y. \]  

(4.11)

Here \( I_X \) and \( I_Y \) are information matrices of the complete and observed data respectively in the true Fisherian sense and \( I_{X|Y} \) is the average of the Fisher information \( J_{\text{obs}} \) over \( Y \). Although (4.11) seems to give better interpretation, it happens that (4.9) is more useful in characterizing the variance of the random perturbation in Stochastic EM. We shall see that in Section 3.

### 4.2 Mean of the Stationary Distribution

Observe that in many of the examples given in Section 3 of Chapter 3, the mean of the stationary distribution is exactly the MLE. Therefore, it is natural to consider the mean of the stationary distribution as an alternative estimator of the parameter \( \theta \). Although generally the exact stationary distribution \( \Psi \) is unknown, it can always be approximated by its empirical version.

To be specific, suppose the Stochastic EM algorithm is terminated after \( M \) iterations. The estimator we propose is the average

\[ \bar{\theta} = \frac{1}{M - m_0} \sum_{m=m_0+1}^{M} \theta^{(m)}. \]  

(4.12)

The integers \( M \) and \( m_0 \) are chosen large enough to ensure that the Markov chain \( \{\theta^{(m)}\} \) is close to its stationary regime for \( m > m_0 \). In fact, the number
of iterations $M$ can always be chosen large enough to ensure that $\hat{\theta}$ approximates the mean, Mean($\Psi_n$) of the distribution $\Psi$ to any given precision. Therefore, with a slight abuse of notation, we denote the estimate in (4.12) by Mean($\Psi_n$).

In practice, it would be preferable to have a point estimate using only the Stochastic EM iterations than having to write an extra set of codes for EM, even when EM is feasible for solving the problem at hand. Therefore, it is desirable for Mean($\Psi_n$) to have optimal properties.

In this section, we develop a result for this estimate: Mean($\Psi_n$) = $\hat{\theta} + O(1/n)$, when the complete data is from an exponential family. Here $\hat{\theta}$ denotes the maximum likelihood estimator.

Recall from Section 1 that when $x = (y, z)$ is distributed as (4.2), the sufficient statistic for $\theta$ is linear in the loglikelihood. This is the situation where EM can be decomposed into two distinct steps, the E-step and the M-step. In fact, this is the situation in which results for EM can be established in an elegant and rigorous manner (Rubin 1991; Wu 1983). To emphasize the multidimensional nature of the parameters and observations, vectors are represented in bold face. Furthermore, when no confusion arises, we write $T$ for $T_{EM}$ for the sake of clarity.

To begin, let $x = (y, z)$ and take logarithm in (4.2):

$$\log f ((y, z) | \theta) = < \theta, t(y, z) > - A(\theta) + \log h(x). \quad (4.13)$$

Suppose that a solution to the likelihood equation (4.13) exists so that the MLE is some function of $t(y, z)$:

$$\hat{\theta}(y, z) = \Lambda (t(y, z)). \quad (4.14)$$

The EM algorithm seeks to solve (4.6) $E_\theta(t) = E_\theta(t | y)$ by first computing for the right hand side,

$$t^{(m)}(y, z) = E_{\theta^{(m)}} [t(y, z) | y], \quad (4.15)$$

which is the E-step, the expectation being taken with respect to $k(z | y, \theta^{(m)})$. Hereafter we write $E [t(y, z) | y, \theta^{(m)}]$ to represent $E_{\theta^{(m)}} [t(y, z) | y]$ for consistency in notation. Then the M-step solves for $\theta$ in

$$E_\theta(t) = t^{(m)}.$$
This derivation indicates that in this situation, the EM operator takes the particular form

\[
\theta^{(m+1)} = T \left( \theta^{(m)} \right) = \Lambda \left( t^{(m)}(y, z) \right) = \Lambda \left( E[t(y, z) \mid y, \theta^{(m)}] \right).
\]

(4.16)

Whereas in Stochastic EM, the estimate of the parameter at the \((m+1)\)-th iteration is given by

\[
\theta^{(m+1)} = \Lambda \left( t(y, z^{(m)}) \right),
\]

with \(z^{(m)}\) simulated from the distribution \(k(z \mid y, \theta^{(m)})\). Denote the stationary distribution of Stochastic EM by \(\Psi_n\). Then

\[
\text{Mean}(\Psi_n) = \int_\Theta \int_Z \hat{\theta}(y, z) k(dz \mid y, \theta) \Psi_n(d\theta) = \int_\Theta E \left[ \hat{\theta}(y, z) \mid y, \theta \right] \Psi_n(d\theta) = \int_\Theta E \left[ \Lambda \left( t(y, z) \right) \mid y, \theta \right] \Psi_n(d\theta).
\]

(4.17)

Suppose that \(E\) and \(\Lambda\) are interchanged in (4.17), then

\[
\int_\Theta E \left[ \Lambda \left( t(y, z) \right) \mid y, \theta \right] \Psi_n(d\theta) = \int_\Theta \Lambda \left( E \left[ t(y, z) \right] \mid y, \theta \right) \Psi_n(d\theta) = \int_\Theta T(\theta) \Psi_n(d\theta).
\]

(4.18)

In order to interchange \(E\) and \(\Lambda\) with an impunity of order \(1/n\), we use the delta-method for moments. The delta method requires boundedness conditions on the function \(\Lambda\) or its derivatives (Bishop, Fienberg and Holland 1975; Lehmann 1988; Oehlert 1992).

Follow a theorem in Lehmann (1988 page 109, theorem 5.1a), we provide the following lemma without proof.

**Lemma 4.1** Let \(X_1, \ldots, X_n\) be independent and identically distributed with \(E(X_1) = \xi\), \(\text{var}(X_1) = \sigma^2\) and finite \(p\)-th moment where \(p \geq 4\). If \(h\) is a function of a real variable whose first \(p\) derivatives exist and the \(p\)-th derivative is bounded, then

\[
E[h(\bar{X})] = h(\xi) + \frac{\sigma^2}{2n} h''(\xi) + R_n,
\]

where \(R_n\) is \(o(1/n)\).
Lemma (4.1) gives conditions under which expectation and \( h(.) \) can be interchanged with an impurity \( O(1/n) \) in a unidimensional case. The general case requires a multivariate version of the lemma and the multivariate Taylor expansion essentially gives us the required tool. Consider \( \Lambda = (\Lambda_1, \ldots, \Lambda_k) \) as a collection of \( k \) functions \( \Lambda_i : \mathbb{R}^k \to \mathbb{R}, \ i = 1, \ldots, k \). The conditions for the delta method to apply in this multivariate setting are

- (L.1) the existence of the first \( p \) derivatives of all variables and the boundedness of the \( p \)-th derivative in each \( \Lambda_i \),

- (L.2) the existence of the first \( p \) moments of the conditional distribution, namely

\[
E(||Z - E(Z)||^p | y\theta) < \infty \quad \text{for all} \theta \in \Theta,
\]

where \( p \geq 4 \) and \( X = t(Y, Z) \). Under these two conditions, it follows from (4.17) that

\[
\text{Mean}(\Psi_n) = \int_{\Theta} T(\theta) \Psi_n(d\theta) + O(1/n).
\]  

(4.19)

Besides an \( O(1/n) \) term, (4.19) asserts that the mean of the stationary distribution is an average of the EM operator over the various \( \theta \) which are distributed according to the stationary distribution.

Suppose further that

- (A.1) For \( \theta \sim \Psi_n, E||\theta - \hat{\theta}||^2 = O(1/n) \),

- (A.2) the EM operator \( T = (T_1, \ldots, T_k) \) has component mappings \( T_j : \mathbb{R}^k \to \mathbb{R}, \ j = 1, \ldots, k \) and \( T_j \) are not too erratic so that \( D^2T_j \), the second order differential matrix of \( T_j \), stays bounded for all \( j \) over \( \theta \).

Then a quadratic Taylor expansion around the MLE \( \theta = \hat{\theta}_n \) yields

\[
\text{Mean}(\Psi_n) = \hat{\theta}_n + \int_{\Theta} DT(\hat{\theta}_n) (\theta - \hat{\theta}_n) \Psi_n(d\theta) + O(1/n),
\]

where \( DT \) denotes the first order differential of \( T \), from which

\[
(I - DT(\hat{\theta}_n)) \text{Mean}(\Psi_n) = (I - DT(\hat{\theta}_n)) \hat{\theta}_n + O(1/n)
\]

follows. Since the MLE \( \theta_n \) exists, it is a fixed point of \( T \) and the spectral radius of \( DT(\hat{\theta}_n) \) is less than one. This implies \( \text{Mean}(\Psi_n) = \hat{\theta}_n + O(1/n) \).

The above exposition can be summarized into the following theorem:
Theorem 4.1 Suppose the complete data are generated from (4.2) and \( \hat{\theta}_n \), the MLE of \( \theta \), has the form (4.14). If (A.1), (A.2), the polynomial bounded \( \Lambda(.) \) assumption (L.1) and the finite moment assumption (L.2) are satisfied, then

\[
\text{Mean}(\Psi_n) = \hat{\theta}_n + O(1/n).
\]

The main reason for such a strong result is that \( \Lambda \) is required to be polynomial bounded. This, however, is not true in general for exponential families under canonical parameterization. The following example illustrates this fact.

Example 4.1. Gamma distribution.

For a gamma distribution with density

\[
\frac{1}{\Gamma(a)b^a} x^{a-1}e^{-x/b}, \quad x > 0, \ a > 0, \ b > 0,
\]

and \( a \) assumed known, the canonical parameter is \( \theta = -1/b \). Also \( A(\theta) = a \log(-\theta) \). By virtue of (4.4) the likelihood equation is

\[
\frac{\alpha}{\bar{\theta}} = -\bar{t},
\]

where \( \bar{t} \) is the mean of the observations. In this case, \( \Lambda(\bar{t}) = -\alpha/\bar{t} \). This function and its derivatives are unbounded. ◊

Although the canonical or natural parameterization gives a simple form of the loglikelihood, in practice it is often not quite used as naturally as its name suggests. For instance, in the above example, \( b \) can be used as the parameter of interest. This leads to a new function \( \Lambda_b(t) = \alpha t \). Its second derivative is bounded.

Therefore, Theorem 4.1 can be extended in the following way: when applying the Stochastic EM algorithm, for a parameter \( \eta = \phi(\theta) \) in (4.1), envision a Markov chain \( \{\eta^{(m)}\} \) running in parallel with the Markov chain \( \{\theta^{(m)}\} \). The Markov chain \( \{\eta^{(m)}\} \) is generated by taking a transformation of \( \theta \). Specifically, the mean of the stationary distribution for \( \eta \), if it exists, can be written in a form similar to (4.17). Using subscript \( \eta \) to denote functions of the new parameter,

\[
\text{Mean}(\Psi_\eta) = \int_{\Phi} \int_{\mathbb{Z}} \bar{\eta}(y,z) k(z \mid y, \eta) \Psi_\eta(d\eta)
\]

55
\[
\int \Phi \left[ \phi(\hat{\theta}(y, z)) \mid y, \eta \right] \Psi_\eta(d\eta) \\
= \int \Phi \left[ (\phi \Lambda)(t(y, z)) \mid y, \eta \right] \Psi_\eta(d\eta).
\]

Regarding \( \phi \Lambda \) as a new function \( \Lambda_\eta \), the argument in the proof of Theorem 4.1 can be applied analogously. The following corollary summarizes the conclusion.

**Corollary 4.1** Let \( \eta \) be a parameter in (4.1), \( \eta = \phi(\theta) \) is a continuous one-to-one function and the MLE of \( \eta \) has the form \( \Lambda_\eta = \phi \Lambda \). If conditions similar to (A.1),(A.2),(L.1),(L.2) are satisfied, then Mean(\( \Psi_\eta \)) = \( \hat{\eta}_n + O(1/n) \).

Note here that condition (A.1) in Theorem 4.1 and Corollary 4.1 is true under fairly general conditions. Section 6 of this chapter presents such a result.

From a Bayesian perspective, the Bayes estimate based on the complete data differ from the MLE by an order \( O(1/n) \). In the incomplete data exponential families setting, suppose that there is a suitable prior distribution \( \pi(\theta) \) to put on \( \theta \) and let \( \pi(\theta \mid y) \) be the posterior distribution after observing \( y \). A proof similar to that in Theorem 4.1 shows that

\[
\hat{\theta}_B(y) = \text{Mean} (\pi(\theta \mid y)) \\
= \int_{\Theta} T(\theta) \pi(\theta \mid y) \, d\theta + O(1/n) \\
= \hat{\theta}_n + O(1/n).
\]

The Bayes estimate averages \( T(\theta) \) over the posterior distribution \( \pi(\theta \mid y) \) while Stochastic EM averages \( T(\theta) \) over the stationary distribution, which is obtained in a completely data-driven fashion. Both estimates subject to deviations of \( O(1/n) \) from the MLE.

The next several sections are devoted to a different approach in studying the stationary distribution of \( \theta^{(m)} \).

### 4.3 Decomposition of Stochastic Difference Equation

In this section we utilize the exponential families structure of the complete data in the hope of simplifying the general stochastic difference equation (3.1)
for Stochastic EM into a more tractable form. The first lemma, Lemma (4.2) is the key to obtaining the kind of simplification we desire.

Using the notations of Section 1, suppose now a sample \( \mathbf{z}_o \) is imputed for the missing value. 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 \( \log k(\mathbf{z} \mid \mathbf{y}) \) by \( \hat{\theta}^k \). In other words,

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

We use \( \hat{\theta}^c(\mathbf{z}_o) \) and \( \hat{\theta}^k(\mathbf{z}_o) \) to emphasize the dependence of both MLE's on \( \mathbf{z}_o \) when necessary. To illustrate the two different MLEs with an example, the readers can refer to Example 3.2. The complete data MLE is \( by + az \) while the conditional MLE is \( \bar{z} \).

For the pseudo-complete sample \( (\mathbf{y}, \mathbf{z}_o) \), we make the following assumption on the sufficient statistics \( t(\mathbf{y}, \mathbf{z}_o) \),

\[
t(\mathbf{y}, \mathbf{z}_o) = E[t(\mathbf{y}, \mathbf{z}) \mid \theta_o, \mathbf{y}] \text{ for some } \theta_o \in \Theta.
\] (4.22)

It is natural at this point to introduce continuity and differentiability assumptions before stating the lemma. Henceforth in this section 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 (D-L-R 1977; Wu 1983).

**Lemma 4.2** Under (4.22), \( T(\hat{\theta}^k(\mathbf{z}_o)) = \hat{\theta}^c(\mathbf{z}_o) \).

**Proof.** Under (4.22), \( t(\mathbf{y}, \mathbf{z}_o) = E[t(\mathbf{y}, \mathbf{z}) \mid \theta_o, \mathbf{y}] \) for some \( \theta_o \). Regard \( t(\mathbf{y}, \mathbf{z}_o) \) as the value of the sufficient statistics obtained during a particular iteration of EM, starting from the guess \( \theta_o \).

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

To prove \( \theta_o = \arg \max_\theta \ell_{\text{cond}}(\theta; \mathbf{z}_o \mid \mathbf{y}) \), first note that when the full data loglikelihood is from an exponential family with sufficient statistics \( t(\mathbf{y}, \mathbf{z}) \),
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; D-L-R 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 \( \theta_o \) that satisfies (4.22)

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

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

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

\[
\frac{\partial}{\partial \theta} \ell_{\text{cond}}(\theta_o; t(y, z_o)) = 0, \tag{4.23}
\]

which implies that \( \theta_o \) 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. \( \diamond \)

**Theorem 4.2** Suppose the complete data \( x = (y, z) \) consists of \( n \) independent random sample drawn from (4.2). 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 iteration of Stochastic EM. If the following conditions holds,

- (A.1) (4.22) is satisfied for a sufficiently large \( m \).
- (A.2) As \( n \to \infty, r/n \to \tau \) where \( 0 < \tau \leq 1 \),
- (A.3) \( T(w) = T(\theta) + DT(\theta)(w - \theta) + o(||w - \theta||), \) as \( w \to \theta \),
where $||.||$ denotes Euclidean norm, then as $n \to \infty$, the distribution of $\theta^{(m+1)} - T(\theta^{(m)})$ tends to the normal distribution

$$N(0, DT(\theta^{(m)})J^{-1}_{\text{cond}}(\theta^{(m)})DT(\theta^{(m)})),$$

where $J_{\text{cond}}$ is defined in (4.9).

Proof. At the $m$-th iteration 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 (4.2), $\theta^{(m+1)} = T(\hat{\theta}^k)$. Under assumption A.2 and using the multivariate delta method of distribution (Bishop, Fienberg and Holland 1975 page 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 $\theta^{(m)}$. ◇

Remark

Condition (A.2) ensures that the information contained in the observed sample is significant as $n \to \infty$. In many hierarchical models such as those discussed in Example 1.2 and 1.3, the ratio $r/n = 1$.

Lemma 4.3 If the regularity conditions

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

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

See D-L-R (1977 Theorem 4) for a proof. The conditions generally hold for regular exponential families.
Corollary 4.2 Under the setting of Theorem 4.2, 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 (4.3)

$$(DT)J_{\text{cond}}(DT)^T = (J_{\text{cond}}J_c^{-1})J_{\text{cond}}^{-1}(J_{\text{cond}}J_c^{-1})$$

$$= J_c^{-1}J_{\text{cond}}J_c^{-1}.$$ 

The result is a consequence of Theorem 4.2. ◊

Remarks:

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

2. Lemma (4.3) gives interpretable quantification of the variance of the stationary distribution. This lemma will be used again in Section 6 in a more general setting.

The next corollary states that the approximate Gaussianity of the kernel is true with other parameterizations.

Corollary 4.3 For any parameter $\eta = \phi(\theta)$, where $\phi(\cdot)$ is continuous and one-to-one, under the same conditions (A.1), (A.2) and (A.3) in Theorem 4.2, as $n \to \infty$, the distribution of $\eta^{(m+1)} - T_n(\eta^{(m)})$ tends to the normal distribution with mean 0 and variance $DT(\eta^{(m)})J_{\text{cond}}(\eta^{(m)})DT(\eta^{(m)}).$

Proof. Note that $T_n = (\phi_nT)$ is the new EM operator. The corollary is a direct result of applying the delta method for distributions. ◊

We have accomplished an important decomposition of

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

The dynamic process underlying Stochastic EM in the exponential families paradigm can be approximated asymptotically by the following equation

$$\theta^{(m+1)} = T(\theta^{(m)}) + \frac{S(\theta^{(m)})}{\sqrt{n}}\epsilon^{(m+1)},$$  \hspace{1cm} (4.24)

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 $\theta$, $\epsilon^{(m+1)}$ is a $k \times 1$ random vector independent
and identically distributed from $\mathcal{N}(0, I)$ for $m = 1, 2, \ldots$, and $T$ is the EM operator. By virtue of Corollary 4.3, $\theta$ is not necessarily the canonical parameter. The functions $T(.)$ and $S(.)$ are always assumed to be continuous and has an appropriate number of continuous derivatives. In addition, the Markov chain generated from (4.24) is generally regarded as aperiodic and irreducible (Chan 1990 page 456). These properties are required in proving ergodicity.

### 4.4 Ergodic Result and Existence of Moments

The Stochastic EM iterations described by (4.24) in an exponential family setting is a non-linear autoregressive process of order one. The following theorem asserts geometric convergence of $\{\theta^{(m)}\}$ to a stationary distribution and finite moment results under regularity conditions.

**Theorem 4.3** For the dynamic process (4.24), 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 $\theta \not\in C$,
- (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 (4.24), the transition kernel has a positive and continuous density. Multiply both sides of (4.24) by their transposes. This gives

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

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

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

(4.26)

61
Thus, under assumption (E1) and (E2), for \( \theta \notin C \),

\[
E\{||\theta^{(m+1)}||^2 \mid \theta^{(m)}\} \leq \rho||\theta^{(m)}||^2 + \frac{K}{n}||\theta^{(m)}||^2 \\
< \rho'||\theta^{(m)}||^2,
\]

for some \( 0 < \rho' < 1 \) when \( n \) is large enough. Use Theorem 3.2 in Chapter 3 with \( g(x) = ||x||^2 \), condition (3.13) is met. From (4.26), 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 (4.26) is \( \text{trace}(ST(\theta)S(\theta))/n \). This term is also uniformly bounded in \( C \) since the trace of the matrix \( ST(\theta)S(\theta) \) 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 (3.14) in Theorem 3.2 is met and geometric ergodicity is proved.

To prove existence of the first two moments, we invoke Theorem 3.3 in Chapter 3 and check on the three conditions that lead to the existence results. Take \( f(\theta) = ||\theta||^2 + 1 \) in order to have \( f \) bounded away from 0. The second condition is met because \( C(= A \text{ in Theorem 3.3}) \) is compact. Finally, the third condition in Theorem 3.3 is satisfied by virtue of (4.28). Therefore we have \( f(1 + ||\theta||^2)\pi(d\theta) < \infty \), implying that the second moment exists. \( \diamond \)

### 4.5 Linear Autoregressive Process

The exact stationary distributions for processes (4.24), even when they exist, are 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 the skeleton \( T \) 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 4.4** Let \( A \) be a \( k \times k \) matrix with \( 0 < ||A|| < 1 \). If

\[
\theta^{(m+1)} = A\theta^{(m)} + b + \epsilon^{(m+1)},
\]

62
where $b, \theta^{(m)} \in R^k$ and $e^{(m+1)}$ is the random perturbation in the $m$-th iteration of Stochastic EM with $e^{(m)} \sim N(0, V)$ independently, then the stationary distribution of $\Psi(\theta)$ is asymptotically normal with mean $\theta = (I - A)^{-1}b$ and variance $\Sigma = \sum_{i=1}^{\infty} A^{i-1}V(A^T)^{i-1}$, where $||V||(1 - ||A||^2) \geq ||\Sigma|| > ||V||$.

Proof. The stationary distribution is given by the solution of the convolution equation of (4.29),

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

(4.30)

where $K(\theta | \tau, y)$ is the Gaussian kernel with mean $A\tau + b$ and variance $V$. It is well known that when $K$ is normal, the solution $\Psi$ is also a normal distribution. In fact (4.29) implies that the solution is given by $\theta = A\theta + b + \epsilon$ or simply $\theta = (I - A)^{-1}(b + \epsilon)$, which is Gaussian. 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 $\Sigma$ is $\exp(it^T\mu)\exp[-(1/2)t^T\Sigma t]$. Our claim is that $\Psi$ is a normal distribution whose mean $\mu$ and variance $\Sigma$ can be determined by substituting the characteristic function of $\Psi$ into (4.30):

$$\exp(it^T\mu - \frac{1}{2}t^T\Sigma t) = \exp(it^T(A\mu - \frac{1}{2}t^T\Sigma A^T t))\exp(it^Tb - \frac{1}{2}t^TVt).$$

(4.31)

Equating the imaginary parts of (4.31) yields $\mu = (I - A)^{-1}b$, which is the MLE $\hat{\theta}$.

Equating the real parts of (4.31) leads to

$$\Sigma - \Sigma AA^T = V.$$  

(4.32)

The formal solution of (4.32) is given by

$$\Sigma = \sum_{i=1}^{\infty} A^{i-1}V(A^T)^{i-1},$$

(4.33)

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 $||A||.||A^T|| < r < 1$ and $||A^{i-1}||.||(A^T)^{i-1}|| \leq Mr^{k-1}$. Therefore, the series in (4.33) converges to a unique solution (Householder 1964).

63
To get bounds for \( ||\Upsilon|| \), first note if \( \Upsilon \) satisfies (4.33), then \( \Upsilon^T \) also satisfies (4.33). Since the solution to (4.33) is unique, \( \Upsilon \) is symmetric. In addition, \( V \) is positive definite implies that \( \Upsilon \) 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 \( \Upsilon \) 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 (4.32) by \( \Gamma^T \) and postmultiply by \( \Gamma \) gives

\[
\Gamma^T \Upsilon \Gamma - \Gamma^T A \Gamma \Gamma^T \Upsilon \Gamma \Gamma^T A^T \Gamma = D. \tag{4.34}
\]

Denote \( \Gamma^T \Upsilon \Gamma \) by \( \Sigma \) and \( \Gamma^T A \Gamma \) by \( B \) and rewrite (4.34) 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 page 216; Mirsky 1955 page 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 \( \rho(\Sigma) = \rho(\Gamma^T \Upsilon \Gamma) = ||\Upsilon||. \) Therefore, from (4.35), \( ||\Upsilon||(1 - ||A||^2) \leq ||V||. \)

To obtain a lower bound note that the matrix \( \Sigma = \Gamma \Upsilon \Gamma^T \) 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 page 510). ◦

In the case where \( \theta \) is unidimensional, the following corollary gives an interpretable solution for the variance of the stationary distribution under the setting of Theorem 4.2.

**Corollary 4.4** If the Stochastic EM iterations are governed by the equation

\[
\theta^{(m+1)} = \rho \theta^{(m)} + b + \epsilon^{(m+1)},
\]

64
where $0 < \rho < 1$, $\epsilon^{(m+1)} \sim \mathcal{N}(0, J_c^{-1} J_{\text{cond}} J_c^{-1})$ independently with $J_c$ and $J_{\text{cond}}$ not dependent on $\theta^{(m)}$, then

1. the stationary distribution $\Psi_n(\theta)$ is $\mathcal{N}(\hat{\theta}, J_c^{-1} \frac{1-\rho^2}{1-\rho^2})$,
2. the variance of $\text{Mean}(\Psi_n)$ can be approximated by $J_c^{-1} \frac{1}{1-\rho}$.

Proof. The MLE $\hat{\theta} = b/(1-\rho)$ is the unique solution of the fixed point equation $\theta = T(\theta) = \rho \theta + b$. We only need to note that $\rho = DT = J_{\text{cond}} J_c^{-1}$ to prove (1). To prove (2), we use $\var(\hat{\theta})$ as an approximation to the variance of the Stochastic EM estimate. Asymptotically, $\var(\hat{\theta}) = J_{\text{obs}}^{-1}$. But $J_{\text{obs}} = J_c - J_{\text{cond}}$. Applying Theorem 4.4 gives the required result. \hfill \Diamond

The next theorem asserts the uniform boundedness of the variance and bias of $\theta^{(m)}$ in situation (4.29). But first, we supply a lemma which is useful.

**Lemma 4.4** Let $B$ be a square $n \times n$ matrix and $\lambda_n$ be its eigenvalue with the largest modulus. Suppose $|\lambda_n| < 1$, then $||y|| > ||By||$ for $y \in \mathbb{R}^n$.

Proof. $||By||^2 = y^T (B^T B) y \leq |\lambda_n|^2 y^T y \leq ||y||^2$. (Ortega 1990) \hfill \Diamond

**Theorem 4.5** Under the same conditions as stated in Theorem 4.4,

1. there exists a constant $B_1 > 0$ such that for all $m$, $E||\theta^{(m)} - E(\theta^{(m)})||^2 < B_1/n$,

2. there exists a constant $B_2 > 0$ such that for all $m$, $E||\theta^{(m)} - E(\theta^{(m)})||^4 < B_2/n^2$,

3. $||E(\theta^{(m)}) - \hat{\theta}||$ is $O(\rho^m)$ for some $0 < |\rho| < 1$,

with all the expectations being conditional on the observed data $y$.

Proof. Let $\mu^{(m)} = E(\theta^{(m)})$ and $V = (1/n)SS^T$. At the $(m+1)$-th iteration,

$$\theta^{(m+1)} = A^{m+1} \theta^{(0)} + (A^m + \ldots + A + I)b + (A^m + \ldots + A + I)S\epsilon/\sqrt{n}.$$ 

Since $E\epsilon = 0$, $\mu^{(m+1)} = A^{(m+1)} \theta^{(0)} + (A^{(m)} + \ldots + I)b$, it follows that

$$||\theta^{(m+1)} - \mu^{(m+1)}||^2 = (1/n)||((I - A)^{-1}(I - A^{m+1})S\epsilon||^2.$$ 

Similarly,

$$||\theta^\infty - \mu^\infty||^2 = (1/n)||((I - A)^{-1}S\epsilon||^2$$
where $\mu^\infty = \hat{\theta}$. Indeed $E[||\theta^\infty - \mu^\infty||^2]$ forms the uniform bound for the sequence $E[||\theta^{(m+1)} - \mu^{(m+1)}||^2]$. Consider $E[||\theta^{(m+1)} - \mu^{(m+1)}||^2 - ||\theta^\infty - \mu^\infty||^2$ and denote the matrix $S^T(X - B^TXB)S$ by $M$ where $X = ((I - A)^{-1})^T(I - A)^{-1}$ and $B = (I - A^m)^{-1}$. The matrix $B$ has spectral norm smaller than 1. Since $X$ is symmetric and positive definite, $P^T XP = D$ where $D$ denotes the diagonal matrix with entries $(\gamma_1, \ldots, \gamma_k)$, and $P$ is orthogonal. Further, let $x \in \mathbb{R}^k$ be a random vector such that $E(xx^T) = I_k$ and $y = Sx, z = P^T y$.

\[
\begin{align*}
x^T M x &= y^T (PDP^T - B^TPDP^T B)y \\
&= z^T Dz - y^T B^TPDP^T By \\
&= z^T Dz - w^T Dw \\
&= \sum_{i=1}^k \gamma_i^2 (z_i^2 - w_i^2) \\
&\geq \gamma_{(1)}^2 (||z||^2 - ||w||^2),
\end{align*}
\]  

(4.35)

where $w = P^T By$. The smallest $\gamma_i$ is denoted by $\gamma_{(1)}$ and $z_i$ and $w_i$ denote the i-th components of the two vectors $z$ and $w$ respectively. Since $z = P^T y$ and $w = P^T By$ and an orthogonal matrix preserves norm, by Lemma 4.4, (4.35) is positive.

Hence $M$ is positive definite. This implies

\[
E[||\theta^\infty - \mu^\infty||^2 - ||\theta^{(m+1)} - \mu^{(m+1)}||^2] = \frac{1}{n} E(x^T M x) = \frac{1}{n} \text{trace}(M) > 0.
\]

Therefore, $E[||\theta^{(m)} - \mu^{(m)}||^2]$ is uniformly bounded by $(1/n) \text{trace}[S^T((I - A)^{-1})^T(I - A)^{-1} S]$, proving (1).

Using a similar argument as above,

\[
\begin{align*}
n^2 E[||\theta^\infty - \mu^\infty||^4 - ||\theta^{(m+1)} - \mu^{(m+1)}||^4] &= (\sum_i \gamma_i^2 z_i^2)^2 - (\sum_i \gamma_i^2 w_i^2)^2 \\
&= (\sum_i \gamma_i^2 (z_i^2 - w_i^2))(\sum_i \gamma_i^2 (z_i^2 + w_i^2)) \\
&> 0.
\end{align*}
\]

proving (2).
For the bias,
\[
||\mu^\infty - \mu^{(m+1)}|| = ||A^{m+1}(I - A)^{-1}b + \theta^{(0)}||
\leq ||A^{m+1}||(I - A)^{-1}b + \theta^{(0)}||
\leq ||A^{m+1}||(I - A)^{-1}b + \theta^{(0)}||,
\]
proving (3). \(\diamondsuit\)

4.6 Non-linear Autoregressive Process

This section presents two theorems concerning the stationary distribution of (4.24) in the exponential families setting where the skeleton is non-linear and the noise is dependent on \(\theta^{(m)}\). In the first theorem, we assume that the Markov process is geometric ergodic and the second moment of the stationary distribution is finite. Note that these conditions can easily be checked using Theorem 4.3 or Theorems 3.2 and 3.3.

**Theorem 4.6** If the Markov process defined by (4.24) is geometric ergodic, its stationary distribution \(\Psi\) has finite second moment and \(T\) has \(\hat{\theta}\) as its unique stable fixed point, then for \(\theta^{(\infty)} \sim \Psi\),
\[
E||\theta^{(\infty)} - \hat{\theta}||^2 = O\left(\frac{1}{n}\right).
\]

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

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

(4.36)
where $M^{(0)} = DT(\xi^{(0)})$, and

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

(4.37)

for some $\Delta^{(0)} = \text{diag}(\delta_i^{(0)})$ where $0 < \delta_i^{(0)} < 1$. Successive substitution of $\theta^{(m)}$ into (4.24) and Taylor expanding as in (4.36) gives

$$
\theta^{(m)} = T^m(\theta^{(0)}) + \frac{1}{\sqrt{n}} \left[ M^{(m-2)} \ldots M^{(0)} S(\theta^{(0)}) \varepsilon^{(1)} \right. \\
+ M^{(m-2)} S(\theta^{(m-2)}) \varepsilon^{(m-1)} + S(\theta^{(m-1)}) \varepsilon^{(m)} 
$$

(4.38)

where $M^{(j)} = DT(\xi^{(j)})$, $j = 0, \ldots, m - 2$, are defined in a similar way as in (4.37). As $m \to \infty$, $T^m(\theta^{(0)}) \to \hat{\theta}$. From (4.38), use the finite second moment assumption gives $E|\theta^{(\infty)} - \hat{\theta}|^2 = O(1/n)$. \diamondsuit

Although equation (4.38) does indicate that $\theta^{(m)}$ is close to $\hat{\theta}$ as $m \to \infty$, it only implies that $E(\theta) = \hat{\theta} + O(1/\sqrt{n})$ instead of the stronger result $E(\theta) = \hat{\theta} + O(1/n)$. However, Theorem 4.6 ascertains condition (A.1) in Theorem 4.1 and Corollary 4.1, where $\text{Mean}(\Psi_n)$ is proved to be $\hat{\theta} + O(1/n)$. We now return to two previous examples and show that their Stochastic EM estimates are close to the MLE.

**Example 2.1. Unknown variance (continued).**

First check the conditions (E.1) and (E.2) in Theorem 4.3 for Example 2.1 in Chapter 2. The form of the function $T$ is

$$
T(\theta) = k_1 \theta^2/(n(\theta + k_2)^2) + \theta k_2/(\theta + k_2),
$$

(4.39)

where $k_1, k_2$ are constants. Clearly there exists a constant $r_1$ such that when $|\theta| > r_1$, $|T(\theta)| < r_2 |\theta|$ for some constant $r_2$. Moreover,

$$
S(\theta) = 2c_1 \theta/((\theta + c_2)^{3/2} + \theta c_2/((\theta + c_2)),
$$

where $c_1, c_2$ are constants. Again, $|S^2(\theta)| \leq r_3 |\theta|^2$ for some constant $r_3$. Hence the two conditions in Theorem 4.3 are satisfied. The chain is geometric ergodic and has finite second moment. Next note that the transition kernel (2.2) is of the form (4.24). Theorem 4.6 gives us the required condition (A.1) in Corollary 4.1. Furthermore, the complete data MLE is a linear
function of the sufficient statistics implying that (L.1) is satisfied. The conditional density is normal implying (L.2) is satisfied. Hence by Corollary 4.1, Mean($\Psi_n$) = $\hat{\theta} + O(1/n)$. ◊

Remark:
Equation (4.39) has two fixed points, $s^2 - \tau^2$ and 0. Equation (4.38) implies that the empirical mean of the stationary distribution should be centered around $s^2 - \tau^2$ or 0 according to whether $s^2 > \tau^2$ or $s^2 \leq \tau^2$. When $s^2 \leq \tau^2$, as $\theta^{(m)} \to 0$, $S(\theta^{(m)}) \to 0$. In practice, Stochastic EM converges to a point mass at 0 under such circumstance. Note that 0 is an absorbing point for the Markov chain $\theta^{(m)}$.

Example 3.7. Grouped multinomial data (Continued).

From (3.34), (3.35),

$$T(\theta) = \frac{k_1}{n} + \frac{k_2(1 - \theta)}{1 + \theta},$$

and

$$S(\theta) = \frac{c_1\theta(1 - \theta)}{(1 + \theta)^2},$$

where $k_1, k_2, c_1$ are constants. Since $0 \leq \theta \leq 1$, the space is compact and stationarity can be achieved at a geometric rate. Actually (E.1) and (E.2) of Theorem 4.3 are satisfied. Hence the chain has a finite second moment and by Theorem 4.6, condition (A.1) of Corollary 4.1 is met. Since $0 \leq \theta \leq 1$, (A.2) on the boundedness of $D^2T$ and finite moment condition of the conditional density (L.2) are met. From (3.33), the complete data MLE is a linear function of the sufficient statistics. Hence by Corollary 4.1, Mean($\Psi$) = $\hat{\theta} + O(1/n)$. ◊

As an extension to the linear case (4.29), the next theorem, Theorem 4.7 deals with the general case in which the kernel is the stochastic difference equation (4.24)

$$\theta^{(m+1)} = T(\theta^{(m)}) + \frac{S(\theta^{(m)})}{\sqrt{n}} e^{(m+1)}$$

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 4.7 are the boundedness 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 4.7, based on a different set of assumptions, gives somewhat more general results than that of Theorem 4.1. It applies to cases where the stochastic equation (4.24) holds, regardless of the model. It also gives the asymptotic distribution of the Markov chain $\{\theta^{(m)}\}$.

**Theorem 4.7** Let the EM operator $T(\theta)$ in (4.24) 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 $DT(\theta) = \frac{\partial T}{\partial \theta}$ 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 $D s_{ij}(\theta) D s_{ij}^T(\theta)$ 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 $\mathcal{N}(0, K(\theta^*))$, where

$$\frac{||S(\theta^*)||^2}{1 - ||A(\theta^*)||^2} \geq ||K(\theta^*)|| > ||S(\theta^*)||^2.$$

**Proof.** To prove (1), observe that in (4.38), $T^{(m)}(\theta^{(0)})$ 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 $m$, (4.38) 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 (4.24):

$$
\theta_i^{(m+1)} = T_i(\theta^{(m)}) + \frac{S_i^T(\theta^{(m)})}{\sqrt{n}} \varepsilon^{(m+1)},
$$

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

$$
\theta_i^{(m+1)} - \hat{\theta}_i = (DT_i)^T U + \frac{1}{2} U^T D^2 T_i(\zeta_i) U + \frac{1}{\sqrt{n}} S_i^T(\theta^{(m)}) \varepsilon^{(m+1)}, \quad (4.40)
$$

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

Condition on $\theta^{(m)} = \theta$ and take expectation on both sides of (4.40), we obtain, with a slight abuse of notation,

$$
E[U_i^{(m+1)} | \theta^{(m)}] = (DT_i)^T U + \frac{1}{2} U^T D^2 T_i(\zeta_i) U, \quad (4.41)
$$

Further take expectation on both sides of (4.41) 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 $b_i^{(m)} = (1/2)E\{(U^{(m)})^T D^2 T_i U^{(m)}\}$,

$$
X^{(m+1)} = AX^{(m)} + b^{(m)}. \quad (4.42)
$$

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), $D^2 T_i(\zeta_i)$ is bounded by $c_i \|\zeta_i\| \leq \|\hat{\theta}\| + \delta \|U\|$ 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 $b$,

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

71
Since $||A^{m+1}X^{(0)}|| \to 0$ and $||\hat{b}||$ is of order $O(1/n)$,

$$\lim_{m \to \infty} EU^{(m+1)} = O(1/n). \quad (4.43)$$

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 (4.43):

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

implying result (2).

Result (3) is a direct consequence of the triangle inequality:

$$||\theta^{(\infty)} - E\theta|| \leq ||\theta^{(\infty)} - \hat{\theta}|| + ||\hat{\theta} - E\theta||.$$

Squaring both sides gives

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

Since $E||\theta^{(\infty)} - \hat{\theta}||^2 = O(1/n)$ by Theorem 4.6 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 (4.40).

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

$$= \sum_{j=1}^{k} \{s_{ij}(\hat{\theta}) + (Ds_{ij}(\xi_{ij}))^T U\}e_j^{(m+1)} \quad (4.44)$$

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

$$||(Ds_{ij}(\xi_{ij}))^T Ds_{ij}(\xi_{ij})|| < c_2 ||\hat{\theta} + \gamma_{ij}U||^2$$

$$\leq \gamma_0 + \gamma_1 ||U|| + \gamma_2 ||U||^2 \quad \text{for some } \gamma_0, \gamma_1, \gamma_2 \ (4.45)$$
Substituting (4.44) in (4.40) gives
\[
\theta_i^{(m+1)} - \hat{\theta}_i = (DT_i)^T U + \frac{1}{2} UT D^2 T_i (\xi_i) U + \frac{1}{\sqrt{n}} \sum_{j=1}^{k} \{ s_{ij}(\hat{\theta}) + (D s_{ij}(\xi_{ij}))^T U \} \epsilon_j^{(m+1)}.
\]

(4.46)

Multiply (4.46) 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 \( U^{(m)} \):
\[
W_i^{(m+1)} = (DT_i)^T W_i^{(m)} + \frac{\sqrt{n}}{2} UT D^2 T_i U + \sum_{j=1}^{k} s_{ij}(\hat{\theta}) \epsilon_j^{(m+1)} + \sum_{j=1}^{k} (D s_{ij} U) \epsilon_j^{(m+1)}.
\]

(4.47)

Identify a new linear order one autoregressive process \( \{Z_i^{(m)}\} \) by defining
\[
Z_i^{(m+1)} = (DT_i)^T Z_i^{(m)} + \sum_{j=1}^{k} s_{ij}^T(\hat{\theta}) \epsilon_j^{(m+1)},
\]

(4.48)

with \( Z_i^{(0)} = W_i^{(0)} \). Expression (4.47) becomes
\[
W_i^{(m+1)} - Z_i^{(m+1)} = (DT_i)^T (W_i^{(m)} - Z_i^{(m)}) + \frac{\sqrt{n}}{2} UT D^2 T_i U + \sum_{j=1}^{k} (D s_{ij} U) \epsilon_j^{(m+1)}.
\]

(4.49)

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

Put together the components \( Z_i^{(m)} \) in (4.48) to form the linear process \( \{Z^{(m)}\} \). In vector form,
\[
Z^{(m+1)} = A(\hat{\theta})Z^{(m)} + S(\hat{\theta}) \epsilon^{(m+1)},
\]

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
\[
E[R^{(m+1)} \mid \theta^{(m)}] = (DT)E[R^{(m)} \mid \theta^{(m)} = \theta] + \frac{\sqrt{n}}{2} p^{(m)},
\]

where \( p^{(m)} \) is a vector with the i-th component equals \( UT D^2 T_i U \). Further take expectation with respect to \( \theta \) and by result in (1),

\[
ER^{(m+1)} = (DT)(ER^{(m)}) + O(1/n),
\]

73
follows. This recursive equation is analogous to (4.42) 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 (4.49) gives

\[
(W_i^{(m+1)} - Z_i^{(m+1)})^2 = [(DT_i)^T(W_i^{(m)} - Z_i^{(m)})]^2 + \frac{n}{4}[U^T D^2 T_i U]^2 + \frac{\sqrt{n}}{2} (DT_i)^T (W_i^{(m)} - Z_i^{(m)}) [U^T D^2 T_i U] \\
+ \sum_{j=1}^{k} (D s_{ij} T_i U) \varepsilon_j^{(m+1)} \varepsilon_j^{(m+1)} \varepsilon_j^{(m+1)} + \text{cross terms involving } \varepsilon_j^{(m+1)}.
\]

(4.50)

We first sum over $i$, then take expectation conditional on $\theta^{(m)} = \theta$ and then unconditional by further taking expectation with respect to $\theta$ conditional on $y$. When we sum (4.50) across $i = 1, \ldots, k$, the first term in the right hand side of (4.50) is exactly $||DT R^{(m)}||^2$, 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

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

(4.51)

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 4.45, $U^T D s_{ij} D^T s_{ij} U = ||D^T s_{ij} U||^2$ is essentially $O(1/n)$. Thus, the sum in (4.51) 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

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

(4.52)
We still require the boundedness of the fourth moment in order to interchange the expectation and limit in (4.52). 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

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

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 (4.52):

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

(4.53)

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

Theorem 4.4 asserts that the stationary distribution of \( \{Z^{(m)}\} \) is given by

\[
Z^{(\infty)}_{(n)} \sim \mathcal{N}(0, K(\hat{\theta})),
\]

(4.54)

putting the subscript \( (n) \) back. The variance matrix \( K(\hat{\theta}) \) equals to \( \sum_{i=1}^{\infty} A_i^{-1} S S^T (A^T)^{i-1} \) with \( A, S \) evaluated at \( \hat{\theta} \). Also, from Theorem 4.4, \( ||S(\hat{\theta})||^2/(1 - ||A||^2) \geq ||K(\hat{\theta})|| > ||S(\hat{\theta})||^2 \). 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 (4.53),

\[
E||W^{(\infty)}_{(n)} - Z^{(\infty)}_{(n)}||^2 \to 0 \quad \text{as} \quad n \to \infty.
\]

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 \( \theta^{(\infty)} \sim \Psi_n \),

\[
\sqrt{n}(\theta^{(\infty)} - \theta^*) \to \mathcal{N}(0, K(\theta^*)).
\]

(4.55)
Remarks:

1. Normality assumption on $\epsilon^{(m)}$ in (4.24) 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 4.7 applies.

Theorem 4.2 states that when the complete data are generated from an exponential family, the Markov kernel of the Stochastic EM iterations 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 4.7 asserts optimal properties of the stationary distribution. The two results can be combined using Lemma 4.3. 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 4.7 hold, then

**Corollary 4.5** \(\theta^{(\infty)} - \theta^* \to \mathcal{N}(0, Q(\theta^*))\) where

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

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

**Proof.** We only need to prove (3). Equation (4.54) in Theorem 4.7 implies

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

By Lemma 4.3, $V(\hat{\theta}) = J_{\text{cond}}J^{-1}_{\theta}J_{\text{cond}}J^{-1}_{\theta}J_{\text{cond}}$ with $J_{\text{cond}} \equiv J_{\text{cond}}(\hat{\theta})$ and $J^{-1}_{\theta} \equiv J^{-1}_{\theta}(\hat{\theta})$. Using the fact that for any square matrices $A$ and $B$, $||BA|| \leq ||A|| ||B||$ and following the proof in Theorem 4.7, result (3) is established.

**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.
Chapter 5

Binary data analysis

Categorical data, and in particular, binary data, carry less information than continuous data and are generally harder to analyze. In hierarchical models such as variance component models, empirical Bayes models and mixture models, the analyses of categorical or binary data are especially challenging. EM has been known to handle such problems with varying degree of success. The major problem is computational — it is often required to perform integrations of non-standard functions in high dimension or integrations of standard functions over a high-dimensional, non-standard space. As observed by Harville and Mee (1984) in their paper on mixed models for ordered categorical data, the evaluation of MLE “will present an insurmountable computational problem, calling, for example, for the integration of pdf $\phi_N(0, V)$ over an $N$-dimensional rectangular region.” Here $\phi_N$ is the $N$-variate Gaussian density function where $N$ is the number of observable data points. See also Searle, Casella and McCulloch (1992) for a discussion of the difficulties encountered in handling variance components of categorical data.

Numerical integrations of the sort described by Harville and Mee (1984) are commonplace in hierarchical models. These integrations make the sometimes conceptually straightforward EM type algorithms computationally prohibitive.

Stochastic EM is particularly appealing in these situations. One approach to handle ordered categorical data is based on threshold models in which it is supposed that the category is determined by the value of a latent or unobservable continuous response variable that follows a particular continuous data model. Suppose that there are $M$ categories. The model assumes that
the k-th category results when the value of the underlying latent variable is located at the k-th of $M$ intervals that partition the real line. By regarding the latent variable as missing data and using stochastically imputed data for these unobserved variables, we can expect inference based on these complete continuous pseudo-data to be much simplified. The difficulty mainly is with simulation. Fortunately, recent advances in simulation technologies such as the Gibbs sampler provide us with the necessary machinery.

Section 1 of this chapter consists of an example illustrating the concept of a latent variable for binary data and an introduction to the Gibbs sampler. Section 2 discusses a variance component example using the Gibbs sampler. A similar but more complicated missing data structure in an empirical Bayes application is analyzed in Section 3.

5.1 Introduction

5.1.1 Latent variable

This subsection illustrates how the concept of a threshold model using an artificially created latent variable can be applied to binary data in a probit regression model.

Suppose that $N$ independent binary random variables $Y_1, \ldots, Y_N$ are observed and each $Y_i$ is distributed Bernoulli($p_i$). Further, suppose for each $Y_i$, there is a vector of known covariates $x_i^T = (x_{i1}, \ldots, x_{ip})$.

A generalized linear model for the response variable $Y_i$ would be to model the expected value of the variable $Y_i$ as a function of the covariates $x_i$.

$$EY_i = p_i = g(x_i^T \beta),$$  \hspace{1cm} (5.1)

where $g(.)$ is the link function (Nelder and McCullagh 1989) and $\beta$ is the parameter of interest. Assume here that $g(.)$ is the normal cumulative density function $\Phi(.)$, which leads to the probit model. We introduce a latent continuous variable $Z_i$ such that

$$Z_i \sim N(x_i^T \beta, 1)$$  \hspace{1cm} (5.2)

and condition on $Z_i$,

$$Y_i = \begin{cases} 
1, & \text{if } Z_i > 0, \\
0, & \text{otherwise.}
\end{cases}$$  \hspace{1cm} (5.3)
Table 5.1: Simulated experiment: comparing EM and Stochastic EM

<table>
<thead>
<tr>
<th></th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>pop.</td>
<td>0.5</td>
<td>-1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>EM</td>
<td>0.701</td>
<td>-0.930</td>
<td>0.397</td>
</tr>
<tr>
<td>St. EM</td>
<td>0.716</td>
<td>-0.940</td>
<td>0.379</td>
</tr>
</tbody>
</table>

This formulation is equivalent to the probit model (5.1) since

$$E[Y] = E[I(Z > 0)] = \text{Prob}(Z > 0) = \Phi(x^T \beta),$$

where $I(.)$ denotes the indicator function. Treating the latent variable $Z_i$ as missing data and invoking Stochastic EM to estimate the parameter $\beta$, the imputation step involves generating $Z_i$ from the conditional distribution $Z_i|Y_i$, which is

$$\mathcal{N}(x_i^T \beta, 1) \text{ truncated at the left by } 0 \text{ when } Y_i = 1,$$

and

$$\mathcal{N}(x_i^T \beta, 1) \text{ truncated at the right by } 0 \text{ when } Y_i = 0. \quad (5.4)$$

After the latent variables $Z_i$ have been filled in with $z_i^{(m)}$, $\beta$ is updated in the M-step,

$$\beta^{(m+1)} = (X^T X)^{-1} X^T z^{(m)}.$$

To illustrate these procedures, we report results using a simulated data set. A 30 by 3 random matrix was first generated. The variables $y_i$ were generated under the model (5.2) and (5.3) with $\beta = (0.5, -1.0, 0.5)$. The Stochastic EM algorithm was started at the point $(-0.5, 2.0, -0.1)$ and was allowed to run for 300 iterations. The average of the last 100 iterations is reported as the Stochastic EM estimate. It is compared to the EM estimate in Table 5.1.

The structure of the above probit model is quite simple and actually it neither requires EM nor Stochastic EM. In hierarchical models where the parameters themselves are modeled as random variates, the missing data structure is more complicated. More sophisticated simulation techniques such as the Gibbs sampler may be required in the imputation step of Stochastic EM. In the next subsection we briefly describe the Gibbs sampler.
5.1.2 Gibbs Sampler

Suppose we are given a joint density \( f(x,y) \) but it is difficult to simulate samples directly from \( f(x,y) \). The Gibbs sampler generates a sample from \( f(x,y) \) by sampling instead from the conditional distributions \( f(x \mid y) \) and \( f(y \mid x) \): start with an initial value \( Y = y_0 \), generate a value \( x_0 \) from the distribution \( f(x \mid Y = y_0) \); alternatively generate \( y_1 \) from the distribution \( f(y \mid X = x_0) \). The Gibbs sequence of random variables

\[
Y_0, X_0, Y_1, X_1, \ldots, Y_k, X_k, \tag{5.5}
\]

can therefore be generated iteratively by the following scheme.

\[
X_j \sim f(x \mid Y = y_j),
\]

\[
Y_{j+1} \sim f(y \mid X = x_j). \tag{5.6}
\]

This scheme of generation of (5.5) from (5.6) is referred to as Gibbs sampling. Under reasonably general conditions, the distribution of \( X_k \) converges to \( f(x) \), the marginal distribution of \( X \) as \( k \to \infty \). In practice, when \( k \) is large enough, we can regard \( (X_k, Y_k) \) as a sample from the joint distribution \( f(x,y) \).

The above idea can easily be generalized to more than two variables. In fact, the usefulness of the Gibbs sampler increases as the dimension of the problem increases. In many applications, functionals of marginal densities of the form

\[
\int \cdots \int g(x)f(x,y_1, \ldots, y_p) \, dy_1 \cdots dy_p. \tag{5.7}
\]

are of interest. The Gibbs sampler allows us to avoid calculating integrals like (5.7). Calculations of the high dimensional integrals can be replaced either by simulation from several tractable multivariate distributions or a series of one-dimensional random variable generations. These ideas are elaborated in the Sections 2 and 3.

The literature on Gibbs sampling is expanding rapidly. For details about its theory, convergence and application, see Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990). Casella and George (1992) gives a concise introduction. For Monte Carlo simulation of non-uniform random deviates, see Devroye (1986).
<table>
<thead>
<tr>
<th>Control Group</th>
<th>Litter</th>
<th>4 days</th>
<th>21 days</th>
<th>Treated Group</th>
<th>Litter</th>
<th>4 days</th>
<th>21 days</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>13</td>
<td>13</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>11</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>13</td>
<td>12</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>5</td>
<td>4</td>
<td>13</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>7</td>
<td>5</td>
<td>14</td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>10</td>
<td>7</td>
<td>15</td>
<td>10</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>16</td>
<td>7</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Weil (1970) data

5.2 Variance Component

The following data set is originally reported by Weil (1970) and later analyzed by Ochi and Prentice (1984), and McCulloch (1994).

The objective of the clinical trial is to test the effect of a chemically treated diet. In the experiment, 16 pregnant rats were given a chemically treated diet while 16 were given a control diet. In each litter, offspring alive after four days were followed to ascertain their survival status at 21 days. The original data are displayed in Table 5.2.

For each rat, the outcome $Y$, whether it survives or not, is therefore binary with

\[
Y = \begin{cases} 
1, & \text{if the rat survives,} \\
0, & \text{otherwise.}
\end{cases}
\]
We use $i = 1, 2$, to index the groups, $j = 1, \ldots, 16$, to index the litters and $k = 1, \ldots, n_j$ to index the individuals. Ochi and Prentice (1984) model the probability of survival of the $k$-th rat in the $j$-th litter within the $i$-th group as a function of the group effect, the litter effect and a random variation at an individual level. Specifically,

$$\text{Prob}(Y_{ijk} = 1 \mid \mu_i, \beta_{ij}) = \Phi(\mu_i + \beta_{ij}).$$

As described in Section 1, we can set up the example as a missing data problem. Introduce a latent variable $Z_{ijk}$ such that

$$Z_{ijk} = \mu_i + \beta_{ij} + \varepsilon_{ijk},$$

where $\varepsilon_{ijk} \sim N(0, 1)$, and

$$Y_{ijk} = 1(Z_{ijk} > 0). \quad (5.8)$$

Here, $\mu_i$ is the treatment effect on the latent scale and $\beta_{ij}$ are the litter effects. Since we are not interested in the survival probability of a particular rat within a particular litter, the litter effect is regarded as random sample from a larger population. That is,

$$\beta_{ij} \sim N(0, \sigma_i^2),$$

independently. Our main interest is in estimating $\mu_i$ and $\sigma_i$ and comparing survival rates for the two different groups. The parameters $\mu_i$ and $\sigma_i^2$ can be estimated by EM via numerical integration. A technical description of its EM implementation is provided by McCulloch (1994).

To implement the Stochastic EM, first note that there are two ‘layers’ of missing data—the first layer being the latent variable $Z_{ijk}$ and the second layer being the random litter effect $\beta_{ij}$. In order to create a pseudo-complete data, we require samples from the joint distribution

$$(\beta_{ij}, z_{ij}) \mid \mu_i^{(m)}, \sigma_i^{(m)}, y. \quad (5.9)$$

Although drawing a sample directly from the joint distribution (5.9) is difficult, it is easy to draw instead from the distribution $\beta_{ij} \mid z, y$ and $z_{ij} \mid \beta, y$. Suppressing the index $i$ for clarity, we propose a Gibbs sampler which simulates, for each group,
<table>
<thead>
<tr>
<th>Treated group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>S.EM</td>
<td>0.942</td>
</tr>
<tr>
<td>MLE</td>
<td>0.946</td>
</tr>
<tr>
<td>O-P</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.3: Analysis of the treated group and the control group. O-P is the estimate provided by the Ochi and Prentice 1984 paper.

\[
\beta_j \mid z, y, \mu, \sigma^2 \sim \mathcal{N}\left(\frac{1}{\sigma^2 + n_j} \left(\frac{1}{\sigma^2} \mu + n_j \bar{z}_j\right), \left(\frac{1}{\sigma^2} + n_j\right)^{-1}\right), \quad (5.10)
\]

and alternatively,

\[
z_{jk} \mid \beta_j, y, \mu, \sigma^2 \sim \mathcal{N}(\mu + \beta_j, 1). \quad (5.11)
\]

truncated at the left by 0 when $y_{jk}$ is 1, truncated at the right by 0 when $y_{jk}$ is 0.

Since the Gibbs sampler is embedded within each Stochastic EM iteration, its implementation has to be efficient for the computation to be feasible. Fortunately, this version of Gibbs sampler reduces the simulations to univariate draws from the truncated normal distribution and can be implemented very efficiently using an acceptance/rejection method (Robert 1992).

After the Gibbs sampler (5.10) – (5.11) is run a sufficient number of times, a sample $\{(\beta_i, Z_i)\}_{i=1}^{N}$ is available. Now with the pseudo-complete sample, the updating of the parameters is straightforward under the normal model (5.8).

In this application, the number of iterations in the Gibbs sampler was 100 and the number of Stochastic EM iterations was 300. The first 200 iterations were used as a burn-in period for the chain to acquire stationarity. Figure 5.1 displays the graphs of Stochastic EM iterations and Table 5.3 summarizes the results for both the treated and the control group.

In both groups, the Stochastic EM estimates Mean($\Psi_n$) agree substantially with the MLE.
Figure 5.1: 300 Stochastic EM iterations for (a) treated group, (b) control group. The numbers on the graphs indicate the iteration.

### 5.3 Empirical Bayes with SAT Data

In this section, an empirical Bayes model concerning cognitive diagnoses in educational testing is discussed. We first provide the background to the problem in Subsection 1. Subsection 2 outlines the model and shows how using Stochastic EM can avoid performing high-dimensional integrations in the estimation of parameters. Finally, Subsection 3 presents the results of an analysis on a SAT data set.

#### 5.3.1 Background and issues

Traditional educational measurement theories such as the Classical Test Theory and Item Response Theory attempt to characterize a student’s abilities in terms of an overall proficiency scale in behavioral domain. For example, a student is compared with other students or his change is gauged from one point in time to another. Measurement models that have been developed are primarily unidimensional in nature. Student abilities are often modeled as a single ability parameter sampled from an universe. Mislevy (1992) points out that “summarizing competence in these terms suits the kinds of low-resource, long-lasting decision it was designed for: sorting, assigning or selecting students into educational activities.” See also Frederiksen and Collins (1989)
for a critical view on traditional tests. Recent advances in cognitive theories, however, indicate that learning is the reorganization and integration of complex tasks. Learning occurs when students begin to perceive relationships, recognize differences, code information into abstract forms, classify and retrieve in a systematic manner and, most importantly, make adjustments to their internal models of the world. These discoveries have made a strong impact on the educational field and motivate a new wave of tests. Tests should no longer be used only for selecting and classifying. They have to become an integral part of instruction. The ultimate goal of tests should be to provide test consumers and instructors information that could lead to the remedy of the students' 'cognitive deficits'.

A number of intelligent tutoring systems were developed in the past decade as a testimony to the change in test theory (e.g. Marshall 1980, Ohlsson and Langley 1985, Gitomer et.al, 1993). They are intended to diagnose cognitive errors made by students through the analysis of test items and responses. Test theories are also developed for analyzing cognitive errors. For example, Haertel (1989) proposes ability lattice for mapping skill structures of achievement test items. Mislevy (1993) uses a probabilistic diagnostic model based on inference network. Tatsuoka (1990) proposes a rule-space approach to construct items suitable for error diagnosis.

Tatsuoka, Birenbaum, Lewis and Sheehan (1992) pursue the issue further by performing a cognitive task analysis on a SAT mathematics test. This test form is used in our subsequent analysis. They code each SAT item into a number of characteristics which they call attributes (Tatsuoka 1990). The attributes are further aggregated and in the final analysis. Fourteen attributes are used in codifying the whole test.

These 14 attributes are:

1. arithmetic
2. elementary algebra
3. advanced algebra
4. geometry and analytic geometry
5. word problems
6. comparison
7. recall and understanding simple computation
8. application of rules and algorithm
9. selection and application of rules and theorems
10. reasoning and logical thinking
11. analytical thinking and cognitive restructuring
12. reading comprehension
13. practical, spontaneous wisdom
14. more than one step in cognitive process

As an example, consider the following test item:

How old was a person exactly 1 year ago if exactly x years ago the person was y years old?
A. y-1
B. y-x-1
C. x-y-1
D. y+x-1
E. y+x

The attributes that are present in this item are attribute 1, attribute 5, attribute 10, and attribute 14. Typically, each item has 3 to 6 attributes. The resulting matrix of 0 and 1 indicating either the absence or presence of attributes across items is called a Q-matrix and will be denoted by X.

Tatsuoka et al. (1992) use an ideal response approach to perform the subsequent analysis. First they reduce the number of students’ possible attribute patterns from $2^{14}$ to a manageable number. Then they determine, based upon expert opinions, a number of ideal responses which correspond to various attribute pattern in a one-to-one manner. The number of ideal response patterns they report is approximately 600. The problem then reduces to that of classification, namely, matching each student’s response to the ‘closest’ ideal response.
5.3.2 Model and inference

Apparently, the approach of Tatsuoka et al. (1992) requires significant effort in the construction of the ideal response patterns and mapping between attribute patterns and response patterns. We propose a statistical model to handle such a data structure, given that the Q-matrix $X$ is provided as a result of task analysis.

There are $J$ independent binary random variables $Y_{i1}, \ldots, Y_{iJ}$, representing the $J$ item responses of the $i$-th individual. Each $Y_{ij}$ is Bernoulli with $P(Y_{ij} = 1) = p_{ij}$, $j = 1, \ldots, J$.

Assume that the $i$-th individual test-taker is characterized by a $p \times 1$ parameter vector $\beta_i$, called the ability parameter vector, signifying his or her proficiency in each of $p$ attributes.

We propose a probit model

$$\text{Prob}(Y_{ij} = 1 \mid \beta_i) = \Phi(x_j^T \beta_i), \quad (5.12)$$

where $x_j = (x_{j1}, \ldots, x_{jp})^T$ is a vector of known binary coded covariates for the $j$-th item. Also define $X = (x_1^T, \ldots, x_J^T)$.

An intercept term may be included to capture the residual difficulty not explained by the attributes. A positive intercept term, for example, indicates that the item is relatively easy. We adhere to the above notation and substitute $(1, x_{j1}, \ldots, x_{jp})$ for $x_j$ when an intercept term is used.

A trivial approach to the problem would be to estimate each individual's parameters independently using maximum likelihood based on model (5.12). However, because any given attribute is present only in a few items, the MLE derived for each $\beta_{ij}$ could be unstable. In many cases, the maximization algorithms would not converge.

In order to stabilize the estimate, one idea is to pool information across students using a parametric empirical Bayes approach (Morris 1983; Efron 1994b). Assume that the individual abilities parameters are sampled from a universe of student abilities, and that the distribution for the ability parameters is $p$-dimensional multivariate normal. That is,

$$\beta_i \sim N_p(\mu, \Sigma).$$

Thomas, Longford and Rolph (1992) propose a normal mixing model for hospital effects on mortality of patients in a medical context and is similar
to the model described in this section. Their data are also binary, and they adopt an empirical Bayes approach. Information is borrowed across hospitals, and, as a result, a more robust estimate is available.

The mean $\mu$ is assumed to be 0 in their application. As in the usual parametric empirical Bayes framework, $\Sigma$ is to be estimated from the data. See also Longford (1994) for a discussion of a similar model.

Following the traditional assumptions made in educational measurement theory, we assume condition on $\beta_i$, the item responses $y_{ij}$ of individual $i$ to be independent.

Therefore, for each individual,

$$P(Y_i = y_i \mid \beta_i) = \prod_j (\Phi(\beta_i^T x_j))^{y_{ij}} (1 - \Phi(\beta_i^T x_j))^{1-y_{ij}}.$$ 

Consequently, the observed likelihood is

$$\prod_{i=1}^{N} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{p} (\Phi(\beta_i^T x_j))^{y_{ij}} (1 - \Phi(\beta_i^T x_j))^{1-y_{ij}} \phi(\beta; \mu, \Sigma) d\beta_1 \cdots d\beta_p,$$

where $\phi(.)$ is the normal density. Since the parameters $\beta_i$ are regarded as a random sample from a universe of test-takers and are not observed, they can be regarded as missing data. Although conceptually EM is quite straightforward, the amount of computation involved can be overwhelming. For the SAT data set described in Subsection 1, the E-step involves a 14-dimensional integral for each student in each parameter and at each iteration. The calculus required in the E-step is similar to that of the NAEP model discussed in Example 2.2 in Section 2.4. A calculation similar to that example shows that if the number of examinees $N$ is in the thousands, the number of 14-dimensional integrations required is of the order 10 million, which is prohibitive.

The Stochastic EM can be applied to avoid performing these numerical integrations. First, we add an extra ‘layer’ of missing values by introducing $N$ latent continuous variables $Z_1, \ldots, Z_N$, each of which is a $J \times 1$ vector. For a fixed $i$, $Z_{ij}$ is distributed $N(x_i^T \beta_i, 1)$. Define $Y_{ij} = 1$ if $Z_{ij} > 0$ and $Y_{ij} = 0$ if $Z_{ij} \leq 0$.

Conceptually, it is useful, although not necessary, to perceive $Z_{ij}$ as a continuous score received by the $i$-th test-taker on the $j$-th item; and we only
observe either 1 or 0 according to whether the score is above or below a threshold.

Implementing Stochastic EM involves the following two steps.

(1) S-step: for the i-th individual, draw a sample from \((\beta^*_i, Z^*_i) \sim \beta, z | y, \mu^{(m)}, \Sigma^{(m)}\).

(2) M-step: Update the parameters \(\mu\) and \(\Sigma\) by their MLEs:

\[
\mu^{(m+1)} = \frac{1}{N} \sum_{i=1}^{N} \beta^*_i,
\]

\[
\Sigma^{(m+1)} = \frac{1}{N} \sum_{i=1}^{N} (\beta^*_i - \mu)(\beta^*_i - \mu)^T.
\]

Note that the S-step in Stochastic EM requires a sample from the joint distribution of \((\beta_i, Z_j)\) whose density is:

\[c \prod_{j=1}^{p} \left\{ \phi(Z_{ij}; x_j^T \beta, 1) y_{ij} I(Z_{ij} > 0) + \phi(Z_{ij}; x_j^T \beta, 1)(1 - y_{ij}) I(Z_{ij} \leq 0) \right\},\]

where \(c\) is the normalizing constant. This density is quite intractable to simulate. However, we can exploit its conditional structure and draw instead from the distribution of each variable conditional on the other, namely \(\beta | z\) and \(z | \beta\). We propose a Gibbs sampler which simulates

\[
\beta_i | y, z, \mu, \Sigma \sim N \left( (X^T X + \Sigma^{-1})^{-1}(\Sigma^{-1} \mu + X^T z), (X^T X + \Sigma^{-1})^{-1} \right);
\]

and, alternatively,

\[
z_{ij} | \beta, y, \mu, \Sigma \sim N(x_j^T \beta, 1),
\]

truncated at the left by 0 when \(y_{ij}\) is 1, and truncated at the right by 0 when \(y_{ij}\) is 0.

The Gibbs sampler (5.13), (5.14) is similar to the one used in Albert and Chib (1993). It is almost a high-dimensional version of (5.8) in the Weil data example.

There are two immediate advantages in using Stochastic EM here. First, it avoids the overwhelming number of high-dimensional integrations which are required in the E-step of EM. Using Stochastic EM also makes the restoration of \(\beta_i\) easy. Once the parameters \(\mu\) and \(\Sigma\) are estimated by the Stochastic EM estimates \(\hat{\mu}, \hat{\Sigma}\) respectively, samples from the conditional distribution
$\beta_i|y, z, \hat{\mu}, \hat{\Sigma}$ are readily available. The important advantage of this Monte Carlo method is that we do not need to write new codes. The codes for conditional distribution can be reused.

### 5.3.3 Results in analyzing SAT data

The data set consists of the responses of 2,000 examinees on one SAT mathematics test, form 8A. The test has 60 items. The characteristics of each item, that is, the attributes, were coded at the Educational Testing Service, Princeton. The number of attributes, $p$, equals 14. Stochastic EM was run for 300 iterations and only the last 100 iterations were used in estimating the parameters. The programs are written in C, and it takes about 5 hours to run on a DEC 5000 machine.

To avoid an identifiability problem, the variance of the first ability parameter is set to 3.0.

The main interests here are two-fold. First, we want to gain insight on the covariance structure of the attributes. Second, we want to obtain information about the ability of each individual examinee on the various attributes.

Since the number of parameters is quite large, the full details of the numerical results are not included here. We only give a concise interpretation of the results. Several low-level cognitive attributes, notably arithmetic skill and knowledge in substantive areas (e.g., algebra, advanced algebra) are found to have a correlation between 0.5 and 0.6. Geometry is the only attribute that stands out as lowly but negatively correlated to most of the other attributes. High-order cognitive skills such as logical thinking and ability to solve problems with more than one step do not appear to correlate with low-order cognitive skills. The variances of the various attributes are comparable. They are listed in Table 5.4. They are about the same magnitude, although there seems to be a greater variability among students in the attribute "selection and application of rules and theorems" (attribute 9).

In order to assess the effect of the length of the Gibbs sampler and the number of Stochastic EM iterations on the outcome of results, further experiments are performed on a smaller simulated data set. We found that the number of Gibbs sampling iterations $G$ need not be very large to get appropriate samples from $(\beta, Z|y, \mu, \Sigma)$. For starting values, we used $Z = 1$ if $Y = 1$ and $Z = -1$ if $Y = 0$. The simulated data set has $p = 3$, $J = 10$ and $N = 200$ and various values of $G = 10, 20, 100, 1000$ were tried. The results
<table>
<thead>
<tr>
<th>attribute</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>variance</td>
<td>3.0</td>
<td>5.8</td>
<td>5.3</td>
<td>5.2</td>
<td>4.7</td>
<td>2.0</td>
<td>3.6</td>
<td>3.3</td>
<td>8.8</td>
<td>4.2</td>
<td>3.4</td>
<td>2.6</td>
<td>3.3</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Table 5.4: Variance components of attributes 1 to 14

turned out to be quite similar. We also tried from 300 to 3000 iterations of Stochastic EM and found no significant difference in the results. For the above SAT data set, \( G = 20 \) was used.

To restore the \( \beta_i \), we use a Monte Carlo approach. Given a particular response pattern \( y_o \), samples \( \beta^{(1)}, \ldots, \beta^{(S)} \) from \( \beta \mid y_o, \hat{\mu}, \hat{\Sigma} \) are drawn. They are obtained by simulating from the joint \( (\beta, Z) \) using the Gibbs sampler (5.13) and (5.14) and keeping only the \( \beta_i \)'s.

The smoothed histograms of \( \beta^{(1)}, \ldots, \beta^{(S)} \) are approximations to the posterior distributions \( \beta \mid y_o, \hat{\mu}, \hat{\Sigma} \). It provides us with more information about the ability parameters of interest than does a single point estimate.

Figure 5.2 shows the posterior distribution of \( \beta_i | y_o, \hat{\mu}, \hat{\Sigma} \) generated by invoking the Gibbs sampler with \( S = 300 \) for two different subjects on two attributes: “arithmetic” and “cognitive restructuring.” The first subject has an overall score of 55%. The second subject, whom we artificially created, gets all items correct. It is interesting to note that the second subject has a much more spread-out distribution.

The empirical Bayes method seems to work better than just treating individual responses separately. For example, when the responses are all 0 or 1, without using information across test-takers, the coefficients \( \beta_i \) would go off to infinity.
Figure 5.2: Distribution of attribute arithmetic (left column) and cognitive restructuring (right column): (a), (b) response pattern has 55 percent correct; (c), (d) response pattern has 100 percent correct
Chapter 6

Standard Error

This chapter proposes a method of obtaining the standard error of the point estimate derived from the Stochastic EM algorithm without having to write new codes. Section 1 describes the method, which is based on the Louis identity (Louis 1982). An example is used to illustrate the method in Section 2. Section 3 presents a comparison between various EM-related methods using a genetic linkage example.

6.1 Monte Carlo Method

Computing the variance estimates is a major question when using EM. Various authors discover various methods to solve the problem. Louis (1982), Meilijson (1989), Meng and Rubin (1991), Carlin (1987) give useful results. Most of them suggest approximations to the variance estimates without requiring the users to write much new codes in addition to those created for EM.

We propose a way of measuring the variability of the estimate Mean($\Psi_n$) that practically requires no new codes other than the Stochastic EM algorithm. The method only uses the codes for the conditional distribution $k(z \mid y, \theta^{(m)})$ to generate Monte Carlo samples and gives approximate solution for the observed information matrix.

Louis (1982) derives an interesting identity relating the curvatures of the observed data loglikelihood surface and the complete data loglikelihood surface:
\[- \frac{\partial^2}{\partial \theta^2} \ell_{\text{obs}}(\theta; y) = E_\theta \{- \frac{\partial^2}{\partial \theta^2} \ell_c(\theta; x) \mid y\} - \text{cov}_\theta \{ \frac{\partial}{\partial \theta} \ell_c(\theta; x) \mid y\}, \quad (6.1)\]

Similar to the Fisher identity (2.9) which relates the gradients of the two loglikelihoods, the key ingredient in the Louis identity (6.1) is the conditional density of the missing data given the observed data and an estimate of the parameter. We propose a conditional parametric bootstrap method (Efron 1994a) in estimating the variance by exploiting (6.1). The main idea is to impute for the missing data \(z^{(1)}, z^{(2)}, \ldots, z^{(M)}\) and compute both terms in the right-hand side of (6.1) by an empirical version:

\[
\frac{1}{M} \sum_{i=1}^{M} - \frac{\partial^2}{\partial \theta^2} \ell_c(\theta; y, z^{(i)}) - \frac{1}{M} \sum_{i=1}^{M} \{ \frac{\partial}{\partial \theta} \ell_c(\theta; y, z^{(i)}) \} \{ \frac{\partial}{\partial \theta} \ell_c(\theta; y, z^{(i)}) \}^T. \quad (6.2)
\]

The parameter \(\theta\) is fixed at Mean(\(\Psi_n\)).

### 6.2 Censored Weibull Data

To illustrate this ideas, we present a simple example using data that are assumed to be generated from a Type I censored Weibull model. For a recent discussion of the MLE in censored Weibull data and applications in reliability see Meeker, Escobar and Hill (1992). The following exposition discusses point estimation first and then explains how the standard error of the estimate can be derived.

Suppose that \(X_i, i = 1, \ldots, n\) is a sample that comes from a Weibull distribution with density

\[f(x; \theta) = \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\alpha}\right)^\beta\right], \quad x > 0,\]

where \(\theta = (\alpha, \beta)\) denotes the scale and shape parameters.

With Type I censoring at a fixed time \(c\), we only observe a censored sample \(\{y_1, \ldots, y_r\}\), \(y_i < c\) for \(i = 1, \ldots, r\). Treat the problem as a typical missing value problem where \(x = \{x_1, \ldots, x_r, x_{r+1}, \ldots, x_n\}\) are the complete data and \(y = \{y_1 = x_1, \ldots, y_r = x_r, y_{r+1} = c, \ldots, y_n = c\}\) are the observed data.
For convenience, denote the \((n-r)\) survival times after \(c\) by \(z = \{z_{r+1}, \ldots, z_n\}\) and \(z_i = w_i + c\) for \(i = r + 1, \ldots, n\) is the data which are not observed. The complete loglikelihood is

\[
\ell_c(\theta|x) = n \log \frac{\beta}{\alpha} + \sum_{i=1}^{n} (\beta - 1) \log \left( \frac{x_i}{\alpha} \right) - \sum_{i=1}^{n} \left( \frac{x_i}{\alpha} \right)^{\beta}.
\]

Finding the MLE involves solving the likelihood equations:

\[
\sum_{i} x_i^\beta \log x_i - \frac{1}{\beta} - \frac{1}{n} \sum_{i} \log x_i = 0, \tag{6.3}
\]

\[
\alpha = \left( \frac{1}{n} \sum_{i} x_i^\beta \right)^{\frac{1}{\beta}}. \tag{6.4}
\]

When the data are censored at a fixed time \(c\), the observed data likelihood is \(\prod_i f(y_i)^{\delta_i} S(c)^{1-\delta_i}\), where \(y_i = \min(x_i, c)\),

\[
\delta_i = \begin{cases} 
1 & \text{if } x_i \leq c, \\
0 & \text{if } x_i > c,
\end{cases}
\]

and \(S(t) = \exp \left[ -(t/\alpha)^\beta \right]\).

Stochastic EM consists of two steps, the stochastic imputation step (S-step) and the maximization step (M-step). We look at these two steps separately.

At the \(m\)-th iterate, given the current estimate \(\theta^{(m)}\), the S-step computes the conditional density \(w\) given \(y\) and \(\theta^{(m)}\):

\[
k(w|c, \theta^{(m)}) = f(c + w|\theta^{(m)})/S(c|\theta^{(m)}) \quad w > 0,
\]

then simulate \((n-r)\) independent \(w_i^{(m)}\), \(i = r+1, \ldots, n\), from the conditional distribution \(k(w|y, \theta^{(m)})\) and use \(z_i^{(m)} = c + w_i^{(m)}\) to ‘fill in’ for the censored data.

The M-step is straightforward. With the pseudo-complete sample \(\{y_1, \ldots, y_r, z_{r+1}^{(m)}, \ldots, z_n^{(m)}\}\) available, we solve the loglikelihood equations (6.3)-(6.4), updating the parameter \(\theta\) to \(\theta^{(m+1)}\). A Newton–Raphson type algorithm is generally good enough to obtain the MLE within ten iterations when started at a position not too far from the maximum. After \(T\) iterations of
the S-step and the M-step, the following equation gives the Stochastic EM point estimate for $\theta$:

$$\text{Mean}(\Psi_n) = \frac{1}{(T - m_0)} \sum_{m=m_0+1}^{T} \theta^{(m)}.$$  

Consider a real data set consisting of 50 points measuring the modulus of rupture of a sample of fir woods. A two parameter Weibull distribution fits the data well. The Cramer–Von Mises statistics is 0.0495 and the Anderson-Darling statistics is 0.346. Both statistics show no significant deviation from the fitted Weibull (Lawless 1982). The histogram of the data and its Weibull fit are shown in Figure 6.1.

The data are then censored at 1.6, corresponding to a censoring rate of 34%. The Stochastic EM and the EM algorithms are both applied to
<table>
<thead>
<tr>
<th>Estimates</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\alpha$ st. error</th>
<th>$\beta$ st. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncensored MLE</td>
<td>1.616</td>
<td>4.651</td>
<td>0.052</td>
<td>0.497</td>
</tr>
<tr>
<td>Mean($\Psi_n$)</td>
<td>1.579</td>
<td>5.431</td>
<td>0.050</td>
<td>0.793</td>
</tr>
<tr>
<td>EM</td>
<td>1.575</td>
<td>5.435</td>
<td>0.051</td>
<td>0.839</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>-</td>
<td>-</td>
<td>0.052</td>
<td>0.790</td>
</tr>
</tbody>
</table>

Table 6.1: Comparison of estimates of various methods for censored Weibull data

the censored data. In this application, $m_o = 200$ and $T = 300$. Table 6.1 summarizes the results for the two estimation methods. The MLE for the original uncensored data are also included as a reference.

Figure 6.2 shows the points visited by the Monte Carlo chain $\theta^{(m)}$ and the location of the various estimates. Note that there may be differences in the estimates for a another run of Stochastic EM with a different random seed. We tried several runs and found that the differences were insignificant.

The standard error for the Stochastic EM estimates is obtained by using the Louis identity (6.1). In order to compute the observed information matrix of the complete data $-\frac{\partial^2}{\partial \theta^2} \ell_{obs}(\theta; y)$ in (6.1), we need the second derivatives of the complete-data loglikelihood:

$$\frac{\partial^2 \ell_c}{\partial \alpha^2} = n \frac{\beta}{\alpha^2} - \left(\frac{\beta + 1}{\alpha}\right) \left(\frac{\beta}{\alpha}\right) \sum_i \left(\frac{x_i}{\alpha}\right)^\beta,$$

(6.5)

$$\frac{\partial^2 \ell_c}{\partial \beta^2} = \frac{-n}{\beta^2} - \sum_i \left[ \log \left(\frac{x_i}{\alpha}\right) \right]^2 \left(\frac{x_i}{\alpha}\right)^\beta,$$

(6.6)

$$\frac{\partial^2 \ell_c}{\partial \alpha \partial \beta} = \frac{-n}{\alpha} + \frac{1}{\alpha} \sum_i \left(\frac{x_i}{\alpha}\right)^\beta + \frac{\beta}{\alpha} \sum_i \left[ \log \left(\frac{x_i}{\alpha}\right) \right] \left(\frac{x_i}{\alpha}\right)^\beta.$$

(6.7)

First, create pseudo-complete samples $\{(x_i^{(1)}), \ldots, (x_i^{(M)})\}$ by appending to the observed data $M$ samples generated from the conditional distribution $k(z \mid y, \hat{\theta})$, with $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ fixed at Mean($\Psi_n$). For each sample, the observed information matrix is obtained using the equations (6.5) – (6.7). By averaging over the $M$ observed information and covariances matrices, we obtain an empirical approximation to the first term $E_{\theta}\{-\frac{\partial^2}{\partial \theta^2} \ell_c(\theta; x) \mid y\}$ on the right hand side of (6.1). The second term $\text{cov}_{\theta}\{\frac{\partial}{\partial \theta} \ell_c(\theta; x) \mid y\}$ in the right hand side of (6.1) is approximated in a similar fashion by first computing $M$ vectors
Figure 6.2: Markov Chain generated by Stochastic EM for fir wood rupture data showing 300 points. U, S, E represent the point estimate derived from uncensored data MLE, Stochastic EM mean and MLE (from EM) respectively.

of the first partial derivatives of the complete data loglikelihood, which are given by the following two equations:

$$\frac{\partial \ell}{\partial \alpha} = n \frac{\beta}{\alpha} + \left( \frac{\beta}{\alpha} \right) \sum_i (\frac{x_i}{\alpha})^\beta,$$  \hspace{1cm} (6.8)  

$$\frac{\partial \ell}{\partial \beta} = \frac{n}{\beta} + \sum_i \log(\frac{x_i}{\alpha}) - \sum_i [\log(\frac{x_i}{\alpha})](\frac{x_i}{\alpha})^\beta.$$

(6.9)  

In this application, $M = 200$. The results are reported in the second row of Table 6.1. The standard errors using the observed information matrix at the MLE is also given in the third row of the same table.

To assess the performances of the estimates of the standard errors, we further conducted a bootstrap estimate of these standard errors by resampling from the empirical distribution. One thousand bootstrap samples are drawn from the original censored data. The MLE $\theta^*_b, b = 1, \ldots, 1,000$ is computed for each bootstrap sample. The bootstrap standard error is the empirical
standard deviation of $\bar{\theta}_b^*$. That is

$$\sqrt{\frac{1}{B} \sum_{b=1}^B (\bar{\theta}_b^* - \bar{\theta}^*)^2}.$$ 

where $\bar{\theta}^* = (1/B) \sum_b \theta_b^*$ and $B = 1,000$. The result is shown in the last row of Table 6.1. The Stochastic EM estimate of standard error based on the Louis identity seems to work quite well relative to the bootstrap standard error. It must be pointed out that in our experiment we found that the standard error provided by the Louis identity depends quite heavily on $M$, the number of imputed samples. When values less than 100 are used, the answers fluctuate. We recommend using at least $M = 200$.

### 6.3 Genetic Linkage

In order to compare the estimate of standard error obtained by our Monte Carlo method and other existing methods, we consider the celebrated genetic linkage example analyzed in D-L-R (1977), Louis (1982), Tanner and Wong (1987) and Meng and Rubin (1991). Segal et al. (1994) adapt the Meng and Rubin (1991) method to the penalized likelihood case. Their result will also be included in our comparison.

The complete data $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)^T$ have a multinomial distribution with cell probabilities $(1/2, \theta/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4)$. The observed data are $Y_{obs} = (Y_1 + Y_2, Y_3 + Y_4, Y_5)^T = (125, 18, 20, 34)^T$. For the complete data, the observed information is

$$-\frac{\partial^2 l_c(\theta)}{\partial \theta^2} = \frac{Y_2 + Y_5}{\theta^2} + \frac{Y_3 + Y_4}{(1 - \theta)^2}, \quad (6.10)$$

while its score function is

$$\frac{\partial l_c(\theta)}{\partial \theta} = \frac{Y_2 + Y_5}{\theta} - \frac{Y_3 + Y_4}{(1 - \theta)}. \quad (6.11)$$

The Stochastic EM algorithm started at $\theta = 0.5$ and run for 300 iterates. Convergence appeared to be fast and we took Mean($\Psi_n$) to be the average of the last 200 iterates. Standard error for Mean($\Psi_n$) was then obtained via the Louis identity with $M = 100$. Table 6.2 shows our result and results
Table 6.2: Genetic linkage: comparison of various methods. ana. = analytic result; boot = bootstrap; MR = Meng and Rubin (1991); St. EM = Stochastic EM; SBJ = Segal et al. (1994).

Obtained by other authors. To further assess the variability of the standard error estimate due to stochastic noise in simulation, we repeated the above Louis procedure 20 times and report the standard error in parenthesis in Table 6.2. In the penalized likelihood case, the penalty used is \(10(\theta - 0.5)^2\), same as the one used in Green (1990) and Segal et al. (1994). Note that analytic results are available for both the penalized and non-penalized case. Since the validity of all methods discussed here rely on the quadratic behavior of the loglikelihood around the MLE, we also include a small bootstrap experiment for comparison. Two thousand bootstrap samples were drawn from the original data. For each bootstrap sample, the MLE was computed. The bootstrap standard error reported in Table 6.2 is the empirical standard deviation of the two thousand MLEs.

Remark:

1. Observe that the method we propose basically requires no new codes. The codes for simulating from the conditional density should have been already created in the Stochastic EM program.

2. Efron (1994a) points out that the methods proposed by Meng and Rubin (1991), Meilijson (1989) and Carlin (1987) are basically delta methods and they often underestimate the variance. The above scheme gives an approximation to the observed information matrix and may also be subject to the same kind of warning.

3. An alternate, possibly better way to obtain a standard error for the estimate is to bootstrap from the observed data. First bootstrap \(B\)
independent samples from the original observed data set. For each bootstrap sample, the Stochastic EM algorithm is run so that we obtain a point estimate. The standard error can be estimated by $\theta^*_B$:

$$\sqrt{\frac{1}{B} \sum_{b=1}^{B} (\theta^*_b - \bar{\theta})^2}.$$ 

Evidently this procedure would be computationally extremely intensive. However, since the bootstrap samples are independent, the Stochastic EM algorithm can be performed on parallel processors. For most practical applications, it seems that using Louis identity described above gives a reliable estimate of standard error.
Chapter 7

Concluding Remarks

In this dissertation, we have introduced the Stochastic EM algorithm and studied properties of its estimator $\text{Mean}(\Psi_n)$, the mean of the stationary distribution. The algorithm is potentially useful as a methodology for statistical inference in problems with complex, missing-data structures.

Our theoretical results show that the behavior of the stationary distribution of the Stochastic EM iterates relies heavily on the EM operator $T$. We assume throughout this dissertation that $T$ is well-behaved, as most statistical applications do, and that only a finite number of fixed points are present. Local properties about fixed points have been exploited with tools such as Taylor expansion. Nevertheless, global properties such as time of exit from a region of attraction of a fixed point have not been investigated. The Stochastic EM algorithm bears a strong resemblance to global optimization algorithms such as simulated annealing (Geman and Geman 1984; Kirkpatrick and Gelatt 1983). Global large deviation results of Stochastic EM should be of importance to statistical problems with multiple maxima or problems where simulated annealing could be applied. Future research will determine whether Stochastic EM is appropriate for such problems.

The Stochastic EM algorithms are computationally intensive. As the SAT data example in the last chapter illustrates, a Gibbs sampler may have to be embedded within each Stochastic EM iteration. This would be the typical case when the missing-data structure is one of high dimension. Even though computer power has been increasing at a tremendous rate, it would still be wise to keep the amount of simulations manageable. Hence, we shall continue this research in search of improvements in the methodology to reduce the
amount of simulations. One feasible solution is to use crude approximations at the burn-in period and gradually increase the accuracy of the approximation to the proper distribution. Another major problem that might arise in applying the Stochastic EM method is that the Markov chain $\theta^{(m)}$ might be trapped at one of the absorbing points, if they exist. Although restarting the algorithm using a new starting point may be a feasible solution, too many restarts could cause significant bias to the final estimator. We are still investigating this problem.

Another potential area of further research is the connection of Stochastic EM with optimization algorithms. EM is essentially a Gauss-Seidel type algorithm in maximizing functions of the form

$$\sum_i \log \int f(\theta, y_i, z_i) dz_i \quad (7.1)$$

with respect to $\theta$, where $y$ is considered fixed. Almost all optimization algorithms, including EM, require the evaluations of derivatives of some kind of the function, or at least the function itself. Amazingly, the Stochastic EM algorithm seeks an approximation to the maximum of (7.1) without having to evaluate derivatives or even the function itself. It maximizes only functions of the form

$$\sum_i \log f(\theta, y_i, z_i^*) dz_i$$

where $z_i^*$ is an imputed value of $z_i$. Such a property may be of significance in optimization when evaluation of an objective function of the kind 7.1 is expensive.

Applications of the Stochastic EM algorithm to various statistical models with a missing-data structure is a rich area of research and should be pursued further. Qian and Titterington (1991) use a variant of the Stochastic EM algorithm in hidden Markov chain modeling and Masson and Pieczynski (1993) apply the Stochastic EM algorithm to satellite image data. But other applications have been few and far between. Hierarchical models should be an important area that some form of the Stochastic EM algorithm can be applied.

Consider the following pharmacokinetic example. In order to investigate how a drug works, the blood concentration of a drug of a number of patients at various time points are taken after administration of the drug. Let $Y_{ij}$ be the concentration of the $i$-th patient at time $t_{ij}$ and let $w_i$ be a vector of
personal characteristics such as age of the i-th patient. Consider the following hierarchical model
\[ y_{ij} = f(t_{ij}, \theta_i) + \varepsilon_{ij}, \]
\[ \theta_i = g(w_i) + \eta_i, \]
where \( \theta_i \) is an unobserved patient vector parameter and \( \varepsilon_{ij}, \eta_i \) are random variates with mean zero and finite variance. The form of the function \( f(.) \) is known but \( g(.) \) is some non-parametric function whose form cannot be analytically expressed. The function \( g(.) \) can be estimated using certain non-parametric regression procedure such as MARS (Friedman 1989) given a training set of \( (w_i, \theta_i) \). The point of interest is \( g(.) \) and the covariance matrix of \( \eta \). The solution to the above problem is not trivial since there is no known \( \theta_i \). The stochastic imputation idea is straightforward though. We outline the proposed procedures here:

1. Start with an initial guess of the function \( g^{(0)} \).
2. Simulate \( \theta^{(1)} \) from \( \theta \mid y, g^{(0)} \).
3. Update \( g(.) \) by using specific non-parametric regression procedure using the pseudo-complete training set \( (w, \theta^{(1)}) \).
4. Iterate till \( g(.) \) converges.

This application is still under investigation by the author. Several other applications which are also under study by this author include clustering of mixed-mode data (Ip 1994a; Everitt and Merette 1988), representation of multivariate binary data (Ip 1994b) and statistical modelling with parameters under constraints.

Like many other Monte Carlo methods, the solution provided by Stochastic EM is always an approximation. But as John Tukey points out, it is often better to get an approximate solution to a real problem than an exact solution to an oversimplified one. The success of the Stochastic EM algorithm will rely on how good an approximate solution it can provide to problems not easily solved by traditional methods. The optimality of the solution inevitably depends on both its global properties (e.g., ability to transverse the 'energy barrier' in optimization problems) and local properties (e.g., locality of the Mean(\( \Psi_n \) )). Confirmation of these properties of the Stochastic EM estimator should be answered through further research on both its theories and applications.
Bibliography


Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire J. Experimental Agriculture*, 1:129–142.