Computational Statistics (2000) 15:243–277

Computational Statistics

© Physica-Verlag 2000

The S-U algorithm for missing data problems

Glen A. Satten[1] and Somnath Datta[2]

[1]MS E-48, National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA
[2] Department of Statistics, University of Georgia, Athens, GA 30602, USA

Summary

We present a new Monte-Carlo method for finding the solution of an estimating equation that can be expressed as the expected value of a 'full data' estimating equation in which the expected value is with respect to the distribution of the missing data given the observed data. Equations such as these arise whenever the E-M algorithm can be used. The algorithm alternates between two steps: an S-step, in which the missing data are simulated, either from the conditional distribution described above or from a more convenient importance sampling distribution, and a U-step, in which parameters are updated using a closed-form expression that does not require a numerical maximization. We present two numerical examples to illustrate the method. Theoretical results are obtained establishing consistency and asymptotic normality of the approximate solution obtained by our method.

Keywords: Estimating equation; missing data problem; simulate and update; maximum likelihood; EM algorithm; Monte-Carlo method

1 Introduction

Monte-Carlo methods have enlarged the class of problems that statisticians can use to analyze data. Although this expansion has had more effect on Bayesian statistics, Monte-Carlo methods have also been used for frequentist problems, especially to obtain maximum likelihood estimates in problems where exact evaluation of the likelihood is intractable. In many cases of interest, the score or estimating function of the equation to be solved can be written as the expected value of some 'full data' score function with respect to the distribution of the 'missing' data given the observed data. We will refer to this class of problems as 'missing data' problems.

Bayesian Monte-Carlo methods are typically sequential in the sense that they do not require pre-specification of a finite Monte-Carlo sample size. For example, a Gibbs sampler can be run 'over the weekend' and the results collected on Monday, without specifying how many Gibbs steps to take. If the results appear unsatisfactory or if the number of Gibbs steps taken is deemed 'too small,' the sampler can be restarted without loss of the pervious calculation.

In frequentist problems, the quantities of interest are typically the solutions to estimating or score equations. A sequential solution to this frequentist Monte-Carlo calculation can be achieved by constructing a stochastic process which has as its limiting distribution a degenerate distribution centered at the solution of the score equations; this stochastic process can be set up to run 'over the weekend' and intermediate results can be examined without loss of prior calculational effort. When samples from this distribution can be generated easily, a sequential solution to the score equations can be obtained using stochastic approximation (Younes, 1988; Satten, Datta and Williamson, 1998). Although Satten (1996) considers an example where samples from the posterior distribution are generated by using a Gibbs sampler and are then used in a stochastic approximation, there are cases where this sampling cannot be carried out easily and hence cannot be approached using standard methods of stochastic approximation. In particular, currently-available stochastic approximation methods cannot be used when replicates must be generated using importance sampling.

The goal of this work is to develop sequential methods for solving 'missing data' score equations which can be used when replicates are generated using a wide variety of sampling techniques including importance sampling. We call this method the S-U algorithm because it consists of repeated alternation of two steps: an S-step, in which the

missing data are simulated, and a U-step, in which the parameters are updated. In the S-step, one can either use the conditional distribution of the missing data given the observed data (because such sampling can sometimes be carried out without knowing the normalization constant for the distribution, which in this case is the marginal distribution of the observed data), or an importance sampling distribution that is easier to sample from. In the U-step, the parameters are updated using a closed-form expression that does not involve any maximization. As a result, the S-U algorithm can be used in those cases in which the E-M approach is most problematic. The sequence of estimators obtained after each U-step form a stochastic process that converges almost surely to the true zero of the estimating equations.

In missing data problems, the most commonly used approach is Monte-Carlo Maximum Likelihood (MCML), (Geyer, 1991, 1992, 1995; Gelfand and Carlin, 1993) in which a Monte-Carlo approximation to the likelihood is made, and then this approximant is maximized with respect to the parameters of interest. Unfortunately, the number of replicate observations used to construct the Monte-Carlo approximant to the likelihood must be fixed in advance; if it is subsequently desired to increase the number of Monte-Carlo replicates, the calculation must be started over without any benefit from the previous computational effort other than an improved estimate of the starting parameters for the optimization. There is no meaningful way to combine several runs of an MCML (such as averaging).

An additional potential inefficiency in MCML arises because the same replicate data sets are examined r times, where r is the number of steps that the optimization software requires to find the MLE of the approximate likelihood. In a sequential method, such as the S-U algorithm described here, the $\sqrt{n}$ convergence rate ensures that the computational effort required to determine each successive decimal increases by two orders of magnitude. As a result, in a sequential method, most of the time the parameters are fairly near their 'true' values. If a linear approximation to the score function for each replicate data set can be constructed quickly, then it is not necessary to revisit each replicate when our estimate of the parameters change. It should be pointed out that if the Hessian matrix for the full data problem is itself difficult to compute, then this advantage may be lost and the extra evaluations of the likelihood used in MCML may actually be faster.

Another disadvantage of MCML is that it can only be applied in problems where there is a likelihood or objective function to maximize. On the other hand, the S-U algorithm described here applies as long as the estimating equations can be written as the expected value of some 'full data' score taken with respect to a distribution of missing data given

observed data. See Satten Datta and Williamson (1998) for an example using Monte-Carlo methods in such an application.

To our knowledge, excepting MCML, only one other Monte-Carlo method has been proposed for frequentist parameter estimation: the Monte-Carlo EM algorithm of Wei and Tanner (1990). In this approach, the integration involved in the E-step of the E-M algorithm (i.e., obtaining the expected value of the log-likelihood of the 'full data' problem with respect to the posterior distribution of 'missing' given observed data) is carried out using a Monte-Carlo integration scheme. However, this does not converge to the MLE but only fluctuates about the true value; worse, the accuracy of the final iteration depends not on the total number of Monte-Carlo iterates generated but only on the number used in the final step. While this method may be useful for obtaining initial guesses for parameter values, it cannot be recommended for a accurate parameter estimation. In any case where the Monte-Carlo EM can be used, the methodolgy we propose can be used in its place. It is also not advisable to estimate MLEs using a Bayesian Monte-Carlo procedure and then finding the mode of the posterior distribution; the stochastic E-M algorithm (Celeux and Diebold, 1985; Diebold and Ip, 1996) is a Bayesian technique which is not considered further here.

In Section 2, we formally define the problem we are considering, and in Section 3 we propose the S-U algorithm when importance sampling is used. In Section 4, we present theoretical results that ensure the convergence of the S-U algorithm and characterize the nature of the Monte-Carlo error (i.e., the error in estimating the true zero of the estimating equations). In there, we consider the S-U algorithm when sampling directly from the conditional distribution of the missing data given the observed data and compare our approach to the method of stochastic approximation. In Section 5 we compare the S-U algorithm with Monte-Carlo maximum likelihood. In Section 6, we apply the S-U algorithm to a random effects model, and in Section 7 we apply the S-U algorithm to semi-Markov models with interval censored data. We discuss our results in Section 8.

2 Estimating Functions for Missing Data Problems

Suppose that for each individual $i = 1, \cdots, N$, Z_i denotes a random vector, and suppose that Z_i is partitioned into components X_i and Y_i. Observations on Y_i are available, but X_i is unobserved or missing. Let z, x, and y denote realizations of the random variables Z, X, and Y. Let

$S_i(z_i \mid \theta) \equiv S_i(x_i, y_i \mid \theta)$ denote the i-th summand of an estimating function for the complete data, and let $F_\theta(\{x_i\} \mid \{y_i\})$ denote the cumulative distribution function (cdf) of the missing data conditional on the observed data. We are interested in a missing-data estimating function of the form

$$\mathbb{S}_T(\{y_i\} \mid \theta) = \sum_{i=1}^{N} \mathbb{S}_i(y_i \mid \theta) = \sum_{i=1}^{N} \int S_i(x_i, y_i \mid \theta)\, dF_\theta(\{x_i\} \mid \{y_i\}) \quad (1)$$

and in estimators $\hat{\theta}$ obtained from solving

$$\mathbb{S}_T(\{y_i\} \mid \hat{\theta}) = 0\,. \quad (2)$$

For example, the score equations for maximum likelihood are of this form if $S(z \mid \theta)$ is the score function. Note that in some problems $S_i(x_i, y_i \mid \theta)$ may depend on values of x and y for other persons in the study as in Satten, Datta and Williamson (1998). However, for simplicity of presentation we assume that $S_i(x_i, y_i \mid \theta)$ depends only on data from the ith individual, and assume further that $F_\theta(\{x_i\} \mid \{y_i\}) = \prod_i F^i_\theta(x_i \mid y_i)$ in which case we may replace $F_\theta(\{x_i\} \mid \{y_i\})$ by $F^i_\theta(x_i \mid y_i)$ in (1). Let $f^i_\theta(x_i \mid y_i)$ denote the probability density (or mass) function of $F^i_\theta(x_i \mid y_i)$.

To obtain asymptotic variance estimators of $\hat{\theta}$, we also require the Jacobian matrix

$$\mathbb{H}_T(\{y_i\} \mid \theta) = \frac{\partial \mathbb{S}_T(\{y_i\} \mid \theta)}{\partial \theta} = \sum_{i=1}^{N} \mathbb{H}_i(y_i \mid \theta)$$

$$= \sum_{i=1}^{N} \int \left\{ H_i(x_i, y_i \mid \theta) + S_i(x_i, y_i \mid \theta) \underset{\sim}{S}_i^{T}(x_i, y_i) \right\} dF^i_\theta(x_i \mid y_i)$$

$$- \sum_{i=1}^{N} \mathbb{S}_i(y_i \mid \theta) \underset{\sim}{\mathbb{S}}_i^{T}(y_i \mid \theta)$$

$$\equiv \sum_{i=1}^{N} \int \mathcal{H}_i(x_i, y_i \mid \theta)\, dF^i_\theta(x_i \mid y_i) - \mathbb{S}_i(y_i \mid \theta) \underset{\sim}{\mathbb{S}}_i^{T}(y_i \mid \theta)\,, \quad (3)$$

where

$$H_i(x_i, y_i \mid \theta) = \frac{\partial S_i(x_i, y_i \mid \theta)}{\partial \theta} ,$$

$$\tilde{S}_i(x_i, y_i \mid \theta) = \frac{\partial f^i_\theta(x_i, y_i)}{\partial \theta} ,$$

and

$$\tilde{\mathbb{S}}_i(y_i \mid \theta) = \frac{\partial \ln f^i_\theta(y_i)}{\partial \theta} = \int \tilde{S}_i(x_i, y_i \mid \theta)\, dF^i_\theta(x_i \mid y_i) .$$

If $S_i(x_i, y_i)$ is the score function for the likelihood $f^i_\theta(x_i, y_i)$, then $\tilde{S}_i(x_i, y_i \mid \theta) = S_i(x_i, y_i \mid \theta)$ and (3) can be rewritten in terms of the Hessian matrix $H_i(x, y \mid \theta)$ and the variance-covariance function of the score function $S(x, y \mid \theta)$ as

$$\mathbb{H}_T(\{y_i\} \mid \theta) = \sum_{i=1}^{N} \int \Big\{ H_i(x_i, y_i \mid \theta) +$$

$$\Big[S_i(x_i, y_i \mid \theta) - \mathbb{S}_i(y_i \mid \theta)\Big]\Big[S_i(x_i, y_i \mid \theta) - \mathbb{S}_i(y_i \mid \theta)\Big]^T \Big\}\, dF^i_\theta(x_i \mid y_i) .$$

In this case, $\mathbb{H}_T(\{y_i\} \mid \theta)$ is the Hessian matrix of the log-likelihood of the observed data and is symmetric; in the general case, $\mathbb{H}_T(\{y_i\} \mid \theta)$ need not be symmetric.

3 The SU Algorithm

In many cases, $F_\theta(x \mid y)$ is difficult to calculate; even when it is not, the integral in (1) may not be easy to evaluate. In these cases, we propose a class of Monte-Carlo procedures to solve (2) that we will refer to as the S-U algorithm. The S-U algorithm consists of two steps: the S step, in which data are simulated; and the U step, in which the parameters are updated. Unlike the M step in the E-M algorithm, the U step does not correspond to a maximization but is chosen in such a way that the series of

approximants to $\hat{\theta}$, which we will denote by θ_j, $j \geq 1$, converges to $\hat{\theta}$. A single step of the S-U algorithm will correspond to one S step followed by one U step.

In this section, we consider Monte-Carlo solution of (2) using importance sampling. Our approach is based on rewriting the summands of the estimating function in (1) as

$$\mathbb{S}_i(y_i \mid \theta) = \int S_i(x_i, y_i \mid \theta) \frac{f^i_\theta(x_i, y_i)}{f^i_\theta(y_i)\, g^i_\theta(x_i \mid y_i)} \, dG^i_\theta(x_i \mid y_i)$$

$$= \frac{1}{f^i_\theta(y_i)} \int S_i(x_i, y_i \mid \theta)\; w^i_\theta(x_i, y_i)\, dG^i_\theta(x_i \mid y_i) \;,$$

where

$$w^i_\theta(x, y) = \frac{f^i_\theta(x_i, y_i)}{g^i_\theta(x_i \mid y_i)} \tag{4}$$

and where

$$f^i_\theta(y) = \int f^i_\theta(x, y)\, dx = \int w^i_\theta(x, y)\, dG^i_\theta(x \mid y)$$

is the marginal probability distribution function of y. The function $G^i_\theta(x \mid y)$ is the cumulative distribution function for a distribution with density (or mass function) $g^i_\theta(x \mid y)$ which is easy to calculate and sample from and that assigns nonzero mass whenever $f^i_\theta(x \mid y) > 0$. Note that the importance sampling distribution $G^i_\theta(x \mid y)$ may depend on θ.

Similarly, the integrals in (3) can be rewritten as

$$\frac{1}{f^i_\theta(y_i)} \int \mathcal{H}_i(x_i, y_i \mid \theta)\, w^i_\theta(x_i, y_i)\, dG^i_\theta(x_i \mid y_i) \;.$$

Having obtained approximants $\theta_1, \cdots, \theta_j$ to $\hat{\theta}$, at the j-th step of the S-U algorithm let x_{ijk}, $k = 1, \cdots, M$ be iid random variables with distribution $G_{\theta_j}(x \mid y_i)$. Let

$$w_{ijk} = w_{\theta_j}(x_{ijk}, y_i) \;,$$

$$S_{ijk} = S_i(x_{ijk}, y_i \mid \theta_j) ,$$

$$\tilde{S}_{ijk} = \tilde{S}_i(x_{ijk}, y_i \mid \theta_j) ,$$

and

$$H_{ijk} = H(x_{ijk}, y_i \mid \theta_j) .$$

Let

$$\hat{w}_{ij} = \frac{1}{J \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} w_{ij'k} \quad ; \tag{5}$$

note that by averaging over all previous j in (5), $\hat{w}_{ij} \to f_{\hat{\theta}}(y_i)$ as $j \to \infty$ if $\theta_j \to \hat{\theta}$. Similarly, define

$$\hat{\mathbb{S}}_{ij} = \frac{\frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k} w_{ij'k}}{\hat{w}_{ij}} , \tag{6a}$$

$$\tilde{\mathbb{S}}_{ij} = \frac{\frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} \tilde{S}_{ij'k} w_{ij'k}}{\hat{w}_{ij}} \tag{6b}$$

and

$$\hat{\mathbb{H}}_{ij} = \frac{\frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} \left\{ H_{ij'k} + S_{ij'k} \tilde{S}_{ij'k}^{T} w_{ij'k} \right\}}{\hat{w}_{ij}} - \hat{\mathbb{S}}_{ij} \tilde{\mathbb{S}}_{ij}^{T} ; \tag{7}$$

then $\hat{\mathbb{S}}_{ij} \to \mathbb{S}_i(y_i \mid \hat{\theta})$ and $\hat{\mathbb{H}}_{ij} \to \mathbb{H}_i(y_i \mid \hat{\theta})$ as $j \to \infty$, provided $\theta_j \to \hat{\theta}$. Finally, define

$$\hat{\mathbb{S}}_j = \sum_{i=1}^{N} \hat{\mathbb{S}}_{ij} \tag{8}$$

and

$$\hat{\mathbb{H}}_j = \sum_{i=1}^{N} \hat{\mathbb{H}}_{ij} \ ,$$

which approximate the score and Jacobian matrix of the observed data, respectively.

Note that, because using importance sampling to estimate $\mathbb{S}_i(y_i|\theta)$ requires a ratio of estimated terms, a mean-zero estimator of $\mathbb{S}_T(\{y_i\} \mid \theta)$ is not available. Hence, currently-existing stochastic approximation methods such as the Robbins-Monro process cannot be used for Monte-Carlo solution of (2). In the traditional stochastic approximation methodology the θ_j's follow a Markov process; in the S-U algorithm, use of all previous θ_j in determining θ_{j+1} allows $\theta_j \to \hat{\theta}$ even though (8) is only an asymptotically unbiased estimator of $\mathbb{S}_T(\{y_i\} \mid \theta)$.

To carry out the U-step, we approximate $\mathbb{S}_T(\{y_i\} \mid \theta)$ near $\hat{\theta}$ by the linear approximant

$$\hat{\mathbb{S}}_j + \hat{\mathbb{H}}_j \, (\theta - \bar{\theta}_j), \tag{9}$$

where

$$\bar{\theta}_j = \frac{1}{j} \sum_{j'=1}^{j} \theta_{j'} \ ;$$

note that if $\theta_j \to \hat{\theta}$, then (9) approaches the first-order Taylor series to $\mathbb{S}(\{y_i\} \mid \theta)$ at $\hat{\theta}$. Selecting θ_{j+1} to be the zero of this linear approximant results in the updating equation.

$$\theta_{j+1} = \bar{\theta}_j - \hat{\mathbb{H}}_j^{-1} \hat{\mathbb{S}}_j \ . \tag{10}$$

In some cases, it may be desirable to choose an importance sampling distribution $G_\theta(\cdot \mid y)$ for which $g_\theta(x, y)$, but not $g_\theta(x \mid y)$, is easily specified. In this case, we can replace (4) by

$$w^i_\theta(x, y) = \frac{f^i_\theta(x_i, y_i)}{g^i_\theta(x_i, y_i)} \equiv \frac{f^i_\theta(x_i, y_i)}{g^i_\theta(x_i \mid y_i)\, g^i_\theta(y_i)} \; ; \tag{11}$$

hence, if $\theta_j \to \hat{\theta}$ then $\hat{w}_{ij} \to f_{\hat{\theta}}(y_i)/g_{\hat{\theta}}(y_i)$ as $j \to \infty$. Because extra factors of $g_\theta(y_i)$ appear in both the numerator and denominators of (6) and (7), they cancel in the limit as $j \to \infty$, so that $\hat{\mathbb{S}}_{ij}$ and $\hat{\mathbb{H}}_{ij}$ have the proper limiting behavior. To handle both cases simultaneously, define

$$\phi^i_\theta(y) = \begin{cases} f^i_\theta(y) & \text{if (4) is used} \\ \dfrac{f^i_\theta(y)}{g^i_\theta(y)} & \text{if (11) is used;} \end{cases} \tag{12}$$

ϕ_θ will be needed to describe the limiting distribution of θ_j in section 4.

To summarize, the j-th step of the S-U algorithm consists of *simulating* M replicate values for the missing data x from $G^i_{\theta_j}(\cdot \mid y_i)$, computing $\hat{\mathbb{S}}_j$ and $\hat{\mathbb{H}}_j$, and then *updating* the parameter vector θ to its new value θ_{j+1} using (10). The S and U steps are alternated until the desired accuracy has been obtained (or until computational resources are exhausted). In Section 4, we show that the S-U algorithm converges almost surely to $\hat{\theta}$, and that the difference between $\hat{\theta}$ and θ_j is asymptotically normally distributed with a variance-covariance matrix which can be easily estimated.

4 Some Properties of The SU Algorithm

4.1 Consistency of Estimation of $\hat{\theta}$

Given that the S-U algorithm converges to some value θ^*, it is easy to show that $\theta^* = \hat{\theta}$ under mild regularity conditions, when (10) is used for the U step. Specifically, suppose that $\hat{\theta}$ is the unique solution of the estimating equation (2), that $\hat{\mathbb{H}}_j$ converges to $\mathbb{H}_T(\{y_i\} \mid \theta^*)$, which has full rank, and that all elements of this matrix are finite. Since convergence of

θ_j to θ^* implies convergence of $\bar{\theta}$ to θ^*, then equation (10) requires that $\hat{\mathbb{S}}_j \to 0$. However, since $\hat{\mathbb{S}}_j \to \mathbb{S}_T(\{y_i\} \mid \theta^*)$, we must have $\theta^* = \hat{\theta}$. Note that, because only asymptotic unbiasedness of $\hat{\mathbb{S}}_j$ is required, consistency of θ_j when (10) is used can be established for the S-U algorithm with importance sampling.

Under appropriate regularity conditions, the convergence of θ_j can be guaranteed provided the starting value θ_1 is close to $\hat{\theta}$ and a sufficiently large replication size M is chosen. These results are summarized in Theorem 1.

Theorem 1. On a set of probability one, for θ_1 sufficiently close to $\hat{\theta}$ and M sufficiently large, $\theta_j \to \hat{\theta}$ as $j \to \infty$.

Regularity conditions and the proof of Theorem 1 are given in the Appendix.

4.2 Asymptotic Normality of θ_j about $\hat{\theta}$

Assuming that θ_j converges to $\hat{\theta}$ fast enough, specifically that $\frac{1}{\sqrt{j}} \sum_{j'=1}^{j} |\theta_{j'} - \hat{\theta}|$ is bounded in probability, we show that $\sqrt{j}\,(\theta_j - \hat{\theta})$ is asymptotically normally distributed with zero mean and variance-covariance matrix Σ, as $j \to \infty$. In addition, we give an estimator $\hat{\Sigma}$ for Σ that can be calculated recursively as the S-U algorithm is being carried out. These results are summarized in Theorem 2.

Theorem 2. Suppose $\sum_{j'=1}^{j} |\theta_{j'} - \hat{\theta}| = O_p(\sqrt{j})$. Under mild regularity conditions,

$$\sqrt{j}\,(\theta_j - \hat{\theta}) \xrightarrow{d} N(0, \Sigma), \tag{13}$$

where

$$\Sigma = \frac{1}{M}\, \mathbb{H}^{-1}(\{y_i\} \mid \hat{\theta}) \cdot \mathbb{V} \cdot \mathbb{H}^{-T}(\{y_i\} \mid \hat{\theta}) \ . \tag{14}$$

The matrix $\mathbb{V}$ has the following specification. Let V^i be the partitioned matrix

$$V^i = \begin{bmatrix} V^i_{11} & V^i_{12} \\ V^{iT}_{12} & V^i_{22} \end{bmatrix} \tag{15}$$

where

$$V^i_{11} = \int S_i(x, y_i \mid \hat{\theta})\, S_i^T(x, y_i \mid \hat{\theta})\, w^i_{\hat{\theta}}(x, y_i)^2\, dG^i_{\hat{\theta}}(x \mid y_i)$$

$$- \phi^i_{\hat{\theta}}(y)^2\, \mathbb{S}(y \mid \hat{\theta})\, \mathbb{S}^T(y \mid \hat{\theta})\,, \tag{16}$$

$$V^i_{12} = \int S_i(x, y_i \mid \hat{\theta})\, w^i_{\hat{\theta}}(x, y_i)^2\, dG^i_{\hat{\theta}}(x \mid y_i) - \phi^i_{\hat{\theta}}(y)^2\, \mathbb{S}(y \mid \hat{\theta}) \quad , \tag{17}$$

and

$$V^i_{22} = \int w^i_{\hat{\theta}}(x, y_i)^2\, dG^i_{\hat{\theta}}(x \mid y_i) - \phi^i_{\hat{\theta}}(y)^2 \quad . \tag{18}$$

Then

$$\mathbb{V} = \sum_{i=1}^{N} \frac{1}{\phi^i_{\hat{\theta}}(y_i)^2} \Big\{ V^i_{11} +$$

$$\mathbb{S}_i(y_i \mid \hat{\theta}) \mathbb{S}_i^T(y_i \mid \hat{\theta}) V^i_{22} - \mathbb{S}_i(y_i \mid \hat{\theta}) V^{iT}_{12} - V^i_{12} \mathbb{S}_i^T(y_i \mid \hat{\theta}) \Big\}\,. \tag{19}$$

The regularity conditions and proof of Theorem 2 are given in the Appendix.

4.3 Recursive estimation of $\mathbb{V}$, $\mathbb{S}_T(\{y_i\} \mid \theta)$, and $\mathbb{H}_T(\{y_i\} \mid \theta)$

An important feature of the S-U algorithm is that the replicate values x_{ijk} can be processed recursively, and do not need to be stored for later use. To estimate $\mathbb{V}$, define $\hat{V}^{ij}$, partitioned as

$$\hat{V}^{ij} = \begin{bmatrix} \hat{V}^{ij}_{11} & \hat{V}^{ij}_{12} \\ \hat{V}^{ijT}_{12} & \hat{V}^{ij}_{22} \end{bmatrix} \,.$$

We take

$$\hat{V}_{11}^{ij} = \frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k} S_{ij'k}^{T} w_{ij'k}^{2} \quad , \tag{20}$$

$$\hat{V}_{12}^{ij} = \frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k} w_{ij'k}^{2} \quad , \tag{21}$$

and

$$\hat{V}_{22}^{ij} = \frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} w_{ij'k}^{2} \quad . \tag{22}$$

In terms of these and previously defined quantities, the matrix $\mathbb{V}$ can be estimated by

$$\hat{\mathbb{V}}_j = \frac{1}{M} \sum_{i=1}^{N} \frac{1}{\tilde{w}_{ij}^{2}} \left\{ \hat{V}_{11}^{ij} + \hat{V}_{22}^{ij} \hat{S}_{ij} \hat{S}_{ij}^{T} - \hat{S}_{ij} \hat{V}_{12}^{ijT} - \hat{V}_{12}^{ij} \hat{S}_{ij}^{T} \right\} \tag{23}$$

When carrying out the S-U algorithm, for each observation i, we store only the values of $\bar{\theta}_j$, $\hat{w}_{ij}$, $\hat{S}_{ij}$, $\hat{\mathbb{H}}_{ij}$, and $\hat{V}^{ij}$ but not the x_{ijk}'s for the current value of j. Then the values of $\hat{w}_{ij+1}$, $\hat{S}_{ij+1}$, $\hat{\mathbb{H}}_{ij+1}$, and $\hat{V}^{ij+1}$ are easily obtained recursively. The matrix $\mathbb{V}$ need only be calculated at each step if it is needed for a sequential stopping rule; otherwise, it can be calculated after the last S-U step.

4.4 Rejection Sampling

In Section 3 we defined the S-U algorithm using importance sampling. In some cases, it is possible to sample from $F_\theta^i(x \mid y)$ even though we cannot determine the normalization of $F_\theta^i(x \mid y)$ or compute the integral (1). For example, dependent samples from $F_\theta^i(x \mid y)$ can be obtained using a Gibbs sampler, and it is sometimes possible to sample directly from $F_\theta^i(x \mid y)$ by using rejection sampling (Zeger and Karim, 1991). (This can always be accomplished if y is a binary or categorical variable, as is the case in the example considered in Section 6).

The results of Section 3 and the theorems above can be used directly, with $G^i_\theta(x \mid y)$ replaced by $F^i_\theta(x \mid y)$. The U step using (10) is unchanged. Typically, $f^i_\theta(y)$ is difficult to calculate, so (11) would be used in this case. Hence, $w_\theta(x, y) \equiv 1$, so that (6) and (7) simplify to

$$\hat{\mathbb{S}}_{ij} = \frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k} , \quad \tilde{\mathbb{S}}_{ij} = \frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} \tilde{S}_{ij'k}$$

and

$$\hat{\mathbb{H}}_{ij} = \frac{1}{j \cdot M} \left\{ \sum_{j'=1}^{j} \sum_{k=1}^{M} H_{ij'k} + S_{ij'k} \tilde{S}^T_{ij'k} \right\} - \hat{\mathbb{S}}_{ij} \tilde{\mathbb{S}}^T_{ij};$$

while V^i_{12} and V^i_{22} are both identically zero. The expression for $\mathbb{V}$ in (19) simplifies to

$$\mathbb{V} = \sum_{i=1}^{N} V^i_{11} ,$$

while the expression for $\hat{\mathbb{V}}_j$ simplifies to

$$\hat{\mathbb{V}}_j = \frac{1}{M} \sum_{i=1}^{N} \left\{ V^{ij}_{11} - \hat{\mathbb{S}}_{ij} \hat{\mathbb{S}}^T_{ij} \right\} . \tag{24}$$

4.5 Connection to Standard Stochastic Approximation

Because

$$\bar{\mathbb{S}}_j = \sum_{i=1}^{N} \frac{1}{M} \sum_{k=1}^{M} S_{ijk}$$

is an unbiased estimator of $\mathbb{S}_T(y_i \mid \theta_j)$ under rejection sampling, stochastic approximation can be used in cases where the S-U algorithm is used with rejection sampling. For example, the modified Robbins-Monro process of Ruppert et al. (1984) would approximate the solution of (2) by writing

$$\theta_{j+1} = \theta_j - \frac{1}{j} \hat{\mathbb{H}}_j^{-1} \cdot \bar{\mathbb{S}}_j \quad . \tag{25}$$

The relation between (10) and (25) among algorithms that make explicit use of estimates of $\mathbb{H}_T(\{y_i\} \mid \theta)$ is similar to the relationship between an adaptive stochastic approximation scheme such as the Ventner process (see e.g. Ruppert 1991) and Wu's (1985, 1986) stochastic approximation scheme in which the j + 1-th approximant to $\widehat{\theta}$ is the zero of least-squares line obtained by fitting the data $(\theta_j, \bar{\mathbb{S}}_j)$. The asymptotic variance-covariance matrix for the Monte-Carlo error using (25) is also given by (14) and (24) (Ruppert et al., 1984).

5 The SU Algorithm and Monte-Carlo Maximum Likelihood

Monte-Carlo maximum likelihood (MCML) is an alternate procedure that can be used to maximize the likelihood of missing data problems. In this section we compare the MCML and S-U algorithm approaches. Because the Monte-Carlo efficiency of equivalent MCML and S-U approaches are nearly equal, choice between the two procedures can be made based on other factors such as the desirability of a sequential procedure.

MCML can be used for missing data problems when the score equations (1) - (2) arise as a maximization of the likelihood

$$L_\theta = \prod_{i=1}^{N} \int f_\theta^i(x, y_i)\, dx \ .$$

In this paper we will only consider MCML procedures that use importance sampling. As in Section 3, an importance sampling distribution $G_0^i(x \mid y_i)$ can be used to evaluate L_θ. In MCML, a single large sample from $G_0^i(\cdot \mid y_i)$ is chosen; let these samples be denoted x_{ik}, $k = 1, \cdots, M_{mcml}$ and $i = 1, \cdots, N$. Then, θ^* is chosen to be the maximizer of

$$\widehat{\ell}_\theta = \operatorname{Log} \widehat{L}_\theta = \sum_{i=1}^{N} \operatorname{Log} \left\{ \frac{1}{M_{mcml}} \sum_{k=1}^{M_{mcml}} \frac{f_\theta^i(x_{ik}, y_i)}{g_0^i(x_{ik}, y_i)} \right\} . \qquad (26)$$

Geyer and Thompson (1992) propose an iterative strategy, choosing $g_0^i(x, y)$ to be $f_{\theta_o}^i(x \mid y)$, where θ_o is a value of θ that is 'close enough' to $\widehat{\theta}$; in this case, Geyer suggests a series of MCML calculations, setting θ_o for the j-th optimization to be the MCMLE obtained from the (j-1)-th. Once a

'good enough' θ_o is found, a large-scale MCML is carried out. Although this approach may seem similar to the S-U algorithm, it differs in a significant way, as discussed below. Geyer (1995) outlines a second strategy of choosing $g^i_\theta(x \mid y)$ to minimize the variance of some function of the $\hat{\theta}$'s; this approach is not considered further here.

A small generalization of standard results for MCML to the case where $G^i_0(x \mid y)$ is an arbitrary function shows that the asymptotic Monte-Carlo error when maximizing (26) is

$$\sqrt{M_{mcml}}\,(\theta^* - \hat{\theta}) \sim N(0, \Sigma_{mcml}) \tag{27}$$

where

$$\Sigma_{mcml} = \mathbb{H}^{-1}(\{y_i\} \mid \hat{\theta}) \cdot \mathbb{V} \cdot \mathbb{H}^{-T}(\{y_i\} \mid \hat{\theta}) \quad ; \tag{28}$$

aside from the factor of $1/M$ in (14), the only difference between (14) and (28) is that $G^i_{\hat{\theta}}(x \mid y)$ is replaced by $G^i_0(x \mid y)$ in the definition of $\mathbb{V}$. If θ_o is nearly $\hat{\theta}$, then asymptotically, the Monte-Carlo error in MCML is nearly the same as that for the S-U algorithm when an equivalent number of random variables is used (M_{mcml} for MCML; $J \cdot M$ for the S-U algorithm). This equivalence of efficiency is interesting given Geyer's (1995) claim that existing one-sample methods such as MCML are inherently more efficient that many sample methods such as the S-U algorithm because use of (26) estimates ℓ_θ for all θ, not just at the current value θ_i. Although this property of global approximation can be an advantage when θ is far from $\hat{\theta}$ or θ^*, after a few steps ℓ_θ and $\mathbb{S}(\{y_i\} \mid \theta)$ are well approximated by quadratic and linear expansions, respectively. As a result, a many-sample method like the S-U algorithm, which constructs an estimate of the score function that converges to $\mathbb{S}(\{y_i\} \mid \theta)$ near $\hat{\theta}$, can have the same efficiency as MCML.

Furthermore, once a good starting value is available, the S-U algorithm may be preferable, because a sequential approach does not require prior specification of the number of random variables to be generated. In addition, the S-U algorithm allows sequential tuning of the parameters in the distributions that generate the random iterates; with MCML, this tuning can only be accomplished by concluding one MCML calculation and restarting with updated parameters. Doing a series of MCML calculations and choosing the MCMLE of the previous calculation as the value of θ_o for the next MCML calculation, as suggested by Geyer and Thompson (1992) gives no credit for the previous calculations; the

Monte-Carlo random error is a function only of the size of the last MCML calculation.

In the S-U approach, the results from the initializing MCML calculation (to obtain starting values) can be incorporated into the final estimate. Suppose M_{mcml} iterates x_{i0k}, $i = 1,\cdots, N$, $k = 1,\cdots, M_{mcml}$ were generated for a MCML calculation using some fixed importance sampling distribution $G_0^i(x \mid y_i)$ and let the resulting estimate of θ be θ_1. Let $w_{i0k} = w_{\theta_1}(x_{i0k}, y_i)$, $S_{i0k} = S_i(x_{i0k}, y_i \mid \theta_1)$ and $H_{i0k}=H_i(x_{i0k}, y_i \mid \theta_1)$, where we use $G^i_{\theta_1}(\cdot \mid y) = G^i_0(\cdot \mid y)$. Then this information can be incorporated into the SU algorithm with importance sampling by modifying equations (5) and (6) to be

$$\hat{w}_{ij} = \frac{1}{j \cdot M + M_{mcml}} \left[\sum_{k=1}^{M_{mcml}} w_{i0k} + \sum_{j'=1}^{j} \sum_{k=1}^{M} w_{ij'k} \right] \tag{29}$$

and

$$\hat{S}_{ij} = \frac{\frac{1}{j \cdot M + M_{mcml}} \left[\sum_{k=1}^{M_{mcml}} S_{i0k} w_{i0k} + \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k}\, w_{ij'k} \right]}{\hat{w}_{ij}}; \tag{30}$$

equation (7) is modified in an analogous way.

6 A Simple Example

To illustrate use of the S-U algorithm in a simple practical example, we consider a logistic model with repeated measurements for the same individual. We use a random effects model in which the individual-specific intercepts are normally distributed. Although this is not strictly speaking a 'missing data' problem, the score equations for the marginal likelihood are of the form (1)-(2).

Assume a binary response y_{it} and a vector explanatory variables d_{it} are observed for each of N individuals at each time $t = 1 \cdot \cdot \cdot, T$. Assume that the first component of d_i is always 1, and that Y_{it} has a logistic distribution

$$\Pr[\, Y_{it} = y_{it} \mid x_i, \beta, d_{it}] = \frac{e^{(x_i+\beta\cdot d_{it})y_{it}}}{1+e^{(x_i+\beta\cdot d_{it})}}$$

where β are parameters and where x_i is an unknown random component. We assume that the x_i's follow a normal distribution with mean zero and variance σ^2. Hence, the marginal distribution of Y_i is

$$\Pr[\, Y_i = y_i \mid \theta, d_i] = \int \prod_{t=1}^{T} \frac{e^{(x_i+\beta\cdot d_{it})\, y_{it}}}{1+e^{(x_i+\beta\cdot d_{it})}} \; \frac{1}{\sigma}\, \phi(\frac{x_i}{\sigma})\, dx_i \;, \qquad (31)$$

where $\phi(\cdot)$ is the probability density function for the standard normal distribution and where $\theta = (\beta, \sigma)$.

The score functions for θ can be written in the form (1), where

$$S_i(x_i, y_i \mid \theta) = \left(\sum_{t=1}^{T} [y_{it} - L(x_i + \beta \cdot d_{it})]\, d_{it}^T \,,\; \frac{1}{\sigma}(\frac{y_i^2}{\sigma^2} - 1) \right)^T$$

and

$$f_\theta(x_i \mid y_i) = \frac{\prod_{t=1}^{T} \frac{e^{(x_i+\beta\cdot d_{it})\, y_{it}}}{1+e^{(x_i+\beta\cdot d_{it})}} \; \frac{1}{\sigma}\, \phi(\frac{x_i}{\sigma})}{\int \prod_{t=1}^{T} \frac{e^{(x+\beta\cdot d_{it})\, y_{it}}}{1+e^{(x+\beta\cdot d_{it})}} \; \frac{1}{\sigma}\, \phi(\frac{x}{\sigma})\, dx} \;, \qquad (32)$$

and where $L(x)$ is the logit function $(1 + e^{-x})^{-1}$. The integral in (31) and (32) cannot be evaluated analytically, and the expected value of the score functions is intractable. In this particular application, numerical evaluation of (32) has been extensively studied (Crouch and Spiegelman, 1990); as a result, we may compare our results with direct maximization of (31).

Table 1: Data from a 2x2 cross-over trial, reproduced from Diggle, Liang and Zeger, Table 8.1

	Responses			
Group	(1,1)	(0,1)	(1,0)	(0,0)
AB	28	0	6	6
BA	18	4	2	9

Data from a 2x2 cross-over trial on cerebrovascular deficiency adapted from Jones and Kenward (1989) and presented by Diggle, Liang and Zeger (1995) can be analyzed using (31); although the 'usual' analysis for a cross-over trial is conditional logistic regression, the random intercept model allows inclusion of data from the large number of concordant responses, which are noninformative in the usual conditional analysis. These data are presented in Table 1. The responses (1,1), (0,1), (1,0) and (0,0) correspond to the four possible values of (y_{i1}, y_{i2}), where 1 and 0 correspond to a normal and an abnormal electrocardiogram, respectively. Group AB received active treatment at period 1 and placebo at period 2, while group BA received placebo at period 1 and active treatment at period 2. We assume each person's probability of a normal cardiogram follows a logistic model with a person-specific intercept and main effects for period and treatment. The intercept is decomposed into a mean intercept and a person-specific random component, which is assumed to be normally distributed with mean zero and variance σ^2, so that the likelihood (31) applies with $T = 2$ and with β having three components.and with β having three components.

We solved the score equations corresponding to (31) by the S-U algorithm using both importance and rejection sampling. For the S-U algorithm with importance sampling, we used $g^i_\theta(\cdot \mid y_i) = \phi(\cdot/\sigma)/\sigma$, the normal distribution with mean zero and variance σ^2. In this case, $G^i_\theta(\cdot \mid y)$ depends only on one of the components of θ, namely σ_j; hence, updated values of σ are used in $G_\theta(\cdot \mid y)$ at each step. For the S-U algorithm with rejection sampling, we sampled directly from (32) as described in Section 4.4. For each sampling method, we used $M = 100$ and took $j = 100{,}000$ S-U steps. The results of these computations, along with the point estimates for the parameters obtained by direct maximization of the likelihood using numerical integration of (31), are summarized in Table 2. The standard errors when importance sampling is used are approximately 50% larger than those when rejection sampling is used. However, the importance sampling method is approximately twice as fast (approximately four random variables must be generated for each one accepted in the rejection method), leading to a tie in computational efficiency between the two methods in this particular case.

To assess the accuracy of the estimation of the Monte-Carlo error, we conducted a small simulation study. For each sampling method, we estimated $\widehat{\theta}$ 250 times, using $M = 100$ and $j = 1{,}000$ S-U steps. Each simulation was started at $\theta = (0, 0, 0, 1)$. A small-scale Monte-Carlo maximum likelihood calculation (with $M_{mcml} = 500$) was used to start the

Table 2: Comparison of exact parameter estimates with estimates obtained from the S-U algorithm, using M = 100 and 100,000 S-U steps. Standard errors of the Monte-Carlo error of the S-U parameter estimates are given in parentheses.

	Exact	S-U Algorithm: importance sampling	S-U Algorithm: rejection sampling
Intercept	4.0816	4.0842	4.0844
		(0.0036)	(0.0025)
Treatment	-1.8629	-1.8640	-1.8637
		(0.0015)	(0.0010)
Period	-1.0375	-1.0382	-1.0380
		(0.0010)	(0.0006)
σ	4.9431	4.9461	4.9450
		(0.0046)	(0.0031)

S-U algorithm; initial values $\beta = 0$ and $\sigma = 1$ were used, and a normaldistribution with mean 0 and variance 10 was used as the importance sampling distribution for the MCML. The score vector and Jacobian matrix obtained at the end of the MCML was incorporated into the S-U algorithm using (29) - (20). Counting the MCML as the first five steps, the S-U algorithm was then run for an additional 995 steps. At the end of each simulation, the Monte-Carlo error was estimated using (14), (20) and (21)-(23) or (24).

Plots of the nominal coverage of the elliptical confidence regions predicted by Theorem 2 compared with their actual coverage are shown in Figures 1 and 2. For each value of α between 0 an 1, we plot the proportion of simulations for which the $100\alpha\%$ confidence region contains the true mle. Each confidence region was computed using the estimated value of Σ obtained in that simulation. Perfect agreement between nominal and actual coverage corresponds to the 45° line. There is good agreement between nominal and actual coverage, indicating for this problem the values of M and j used are large enough that the asymptotic arguments leading to Theorem 2 are valid.

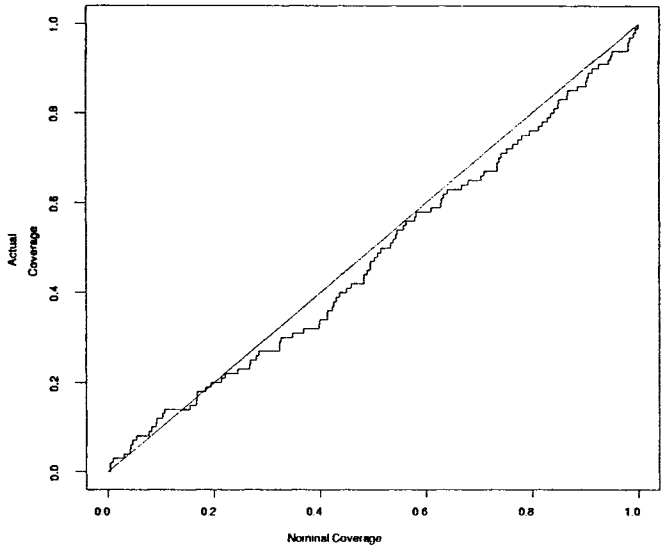

Figure 1 Plot of coverage of $100\alpha\%$ confidence regions as a function of α, for the S-U algorithm using importance sampling, based on results from 250 analyses of the cross-over trial data.

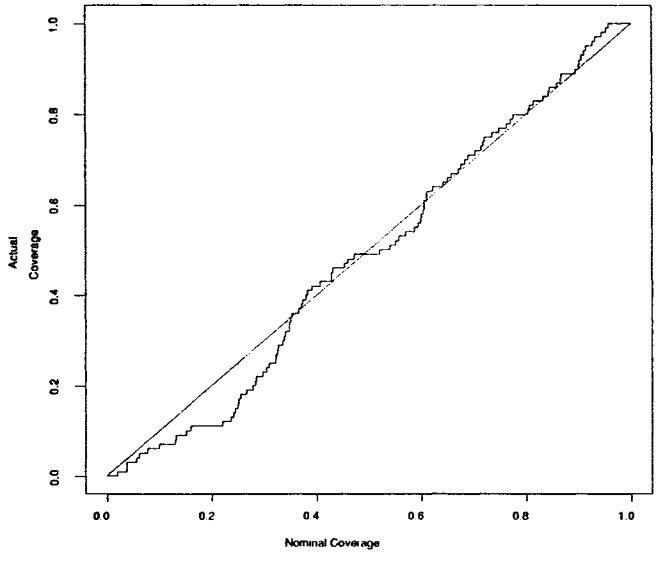

Figure 2 Plot of coverage of $100\alpha\%$ confidence regions as a function of α, for the S-U algorithm using rejection sampling, based on results from 250 analyses of the cross-over trial data.

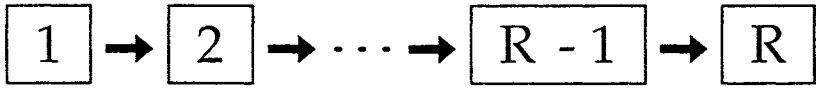

Figure 3. Network of stages for multistage model considered in Section 7. All individuals enter the network in stage 1 at time 0, and move through successive stages until reaching stage R. The ith person waits in stage r for time τ_{ir} and the values of τ_{ir} are independent.

7 A Complex Example

To show the utility of the S-U algorithm in a problem which cannot be addressed by any other sequential Monte Carlo approach, we consider estimation of the waiting time distributions in interval censored chain-of-events data (Sternberg and Satten, 1999). Specifically, consider the multistage model of Figure 3 in which the ith person enters stage 1 at time 0, then enters stage $r+1$ at time $t_{ir}, r = 1, 2, \cdots, R-1$ with $t_{i1} \leq \cdots \leq t_{iR-1}$. We denote the waiting times in each stage by $\tau_{ir} = t_{ir} - t_{ir-1}$ with the convention $t_{i0} \equiv 0$. We assume the transition times follow a semi-Markov process, i.e. the τ_{ir} are independently distributed with probability density function w_θ^r.

We further assume that the t_{ij}'s and τ_{ij}'s are not known exactly, and that the data comprise a series of observation times and stage occupations (*i.e.*, person 1 in stage 1 at time 0, next seen in stage 4 at time 6). Additionally, persons may be censored (*i.e.*, lost to follow-up before reaching stage R).

Our notation loosely follows that of Sternberg and Satten (1999), who considered discrete-time nonparametric models for this type of data. Because it is difficult to test hypotheses using these models, we consider here parametric estimation of the waiting time distributions. Note that knowledge of the t_{ij}'s is equivalent to knowledge of the τ_{ij}'s and *vice versa*. The contribution to the likelihood from the ith person is

$$L_i = \int_{\ell_{i1}}^{u_{i1}} dt_{i1} \cdots \int_{\ell_{in_i}}^{u_{in_i}} dt_{in_i} \prod_{r=1}^{n_i} w_\theta^r(t_{ij} - t_{ij-1})$$

where $[\ell_{ir}, u_{ir})$ is an interval known to contain the (unknown) transition time t_{ir} and $n_i \leq R-1$ is the number of transitions the ith person contributes information to. We will let y_i denote $\{\ell_{i1}, u_{i1}, \cdots, \ell_{in_i}, u_{in_i}\}$ and x_i denote $\{t_{i1}, \cdots, t_{in_i}\}$.

Direct evaluation of the likelihood by numerical evaluation of the n_i-fold integral is unattractive when n_i is large (≥ 3). An E-M approach is also untractable, as the distribution of transition times given the observed data (the ℓ_{ir}'s and u_{ir}'s) requires knowledge of L_i since

$$f_\theta^i(x_i \mid y_i) = \frac{1}{L_i} \prod_{j=1}^{n_i} w_\theta^r(t_{ir} - t_{ir-1})\, I[\ell_{ir} \leq t_{ir} < u_{ir}] \quad . \tag{33}$$

Rejection sampling is computationally simple, but can be very inefficient. A rejection sampling scheme would generate τ_{ir} from w_θ^r and then accept if the resulting transition times t_{ir} satisfy $t_{ir} \in [\ell_{ir}, u_{ir})$ for each $r \leq n_i$. Because an entire new sequence of τ_{ir} values must be calculated if even one of the t_{ir}'s fails to fall in its censoring interval, we can see that rejection sampling can lead to a very high proportion of rejected samples. (Note also that rejection sampling becomes *less* efficient as knowledge of t_{ir} increases, *i.e.* as the intervals $[\ell_{ir}, u_{ir})$ become narrower).

Fortunately, a simple importance sampling scheme can be developed for this type of data. To motivate, consider a simple example with a four-stage model and a person seen twice: in stage 1 at time 0 and stage 3 at time 5. Let W_θ^r and $\overline{W}_\theta^r$ denote the cumulative distribution and survival function for the waiting time distribution in stage r. Sample τ_1 from $f_\theta^1(\tau) I(\tau \leq 5)/W_\theta^1(5)$. Then, sample τ_2 from $f_\theta^2(\tau) I(\tau \leq 5 - t_1)/W_\theta^2(5 - t_1)$. Finally, sample τ_3 from $f_3(\tau) I(5 - t_2 \leq \tau)/\overline{W}_\theta^3(5 - t_2)$. In the most general case, we sample τ_{ir} from

$$d_\theta^r(\tau_{ir} \mid \tau_{i1}, \cdots, \tau_{ir-1}, y_i) = \frac{w_\theta^r(\tau) I\left([\ell_{ir} - t_{ir-1}]^+ \leq \tau \leq u_{ir} - t_{ir-1}\right)}{F_\theta^r(u_{ir} - t_{ir-1}) - F_\theta^r([\ell_{ir} - t_{ir-1}]^+)} \tag{34}$$

where $[x]^+ = x$ if $x > 0$ and 0 otherwise, and where we adopt the convention that individuals censored in stage r have $u_{ir} = \infty$. Then,

$$g^i_\theta(x_i \mid y_i) = \prod_{r=1}^{n_i} d^r_\theta(\tau_{ir} \mid \tau_{i1}, \cdots, \tau_{ir-1}, y_i). \tag{35}$$

Because all previously-proposed recursive schemes require sampling from (33), only the S-U algorithm can utilize this simple importance sampling scheme. While MCML can also be used to solve this problem, it is not recursive and suffers from the inefficiencies described in Section 5.

To demonstrate the ability of the S-U algorithm to solve this problem, we generated data from a four-stage semi-Markov model. The waiting time in stage r followed a Weibull distribution $\overline{W}^r_\theta = \exp\{-a_r t^{b_r}\}$ with scale parameter $a_r = 1/2$ and shape parameter $b_r = r$, $1 \le r \le 3$. Letting $\alpha_r = \ln(a_r)$ and $\beta_r = \ln(b_r)$ we have $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3) \approx$ (-0.693, 0.0, -0.693, 0.693, -0.693, 1.099). Each person was observed 4 times, where the time between successive observations was a uniform[0,1] random variable.

We generated a single dataset with 100 observations using this scheme, and analyzed it 200 times using the S-U algorithm. The value of θ used in the importance sampling distribution (34)-(35) was updated at each S-U step. We stopped the algorithm the first time the largest diagonal element of Σ/j was less than 10^{-6}. We used a block size of M=1000 and each analysis was started at the value obtained by a small (M=1000) MCML calculation. The weights, score and hessian matrix from the MCML were used as initial values for the S-U iterations as in equations (29) and (20).

The results of these calculations are summarized in Table 3. Recall that these simulations all use the same single realization of the semi-Markov process with 100 observations, and hence the average θ_j does not converge to the parameter values used to generate the data, but rather to the mle $\hat{\theta}$ for this particular realization. Because a sequential stopping rule was used, each simulation used a different number of steps j. The mean j was 1129, with range 1112 to 1199. Note in Table 3 that $\sqrt{\Sigma_{rr}/j}$, the estimated standard error of the rth component of $(\theta_j - \hat{\theta})$, is close to its empirical standard error. Note also that there is fairly little variability in $\mathbb{H}^{-1}_{rr}$, the estimated variance of $\hat{\theta}_r$. To show the coverage of the stochastic confidence intervals, a plot of nominal vs. actual coverage of $100\alpha\%$ confidence regions is given in Figure 4. As we have no

independent way to calculate $\widehat{\theta}$ in this example, we used the average θ_j over all simulations as the true mle in calculating Figure 4.

8 Discussion

Monte-Carlo methods have proven to be useful in statistical calculations that are computationally challenging or otherwise prohibitive (Tanner, 1993). In this paper, we have proposed a novel Monte-Carlo method called the S-U algorithm for solving estimating (score) equations such as those that arise in likelihood-based inference for missing data problems or random effects models.

The S-U algorithm was designed to be useful in those cases when the E-M algorithm is most difficult to use. A closed-form expression for the U-step removes the need for a numerical maximization. By using importance sampling, the S-U algorithm makes it unnecessary to compute the

Table 3: Summary of 200 analyses of a single realization of a four stage semi-Markov process using the S-U algorithm.

	Stage 1		Stage 2		Stage 3	
	$\alpha_1=\theta_1$	$\beta_1=\theta_2$	$\alpha_2=\theta_3$	$\beta_2=\theta_4$	$\alpha_3=\theta_5$	$\beta_3=\theta_6$
θ_j						
Mean	-0.6334	0.0154	-0.7631	0.5145	-0.6789	1.2849
Min	-0.6336	0.0148	-0.7635	0.5136	-0.6810	1.2819
Max	-0.6333	0.0158	-0.7627	0.5155	-0.6767	1.2878
S.E. x 10^3	0.0531	0.1759	0.1511	0.4084	0.7924	1.0230
$\sqrt{\Sigma_{rr}/j}$ x 10^3						
Mean	0.0573	0.1928	0.1586	0.4297	0.7128	0.9998
Min	0.0563	0.1895	0.1559	0.4205	0.7040	0.9992
Max	0.0584	0.1968	0.1618	0.4378	0.7253	1.0000
$\mathbb{H}_{rr}^{-1}$						
Mean	0.0195	0.0173	0.0441	0.0405	0.1879	0.1328
Min	0.0195	0.0172	0.0441	0.0402	0.1874	0.1308
Max	0.0195	0.0174	0.0442	0.0408	0.1884	0.1345

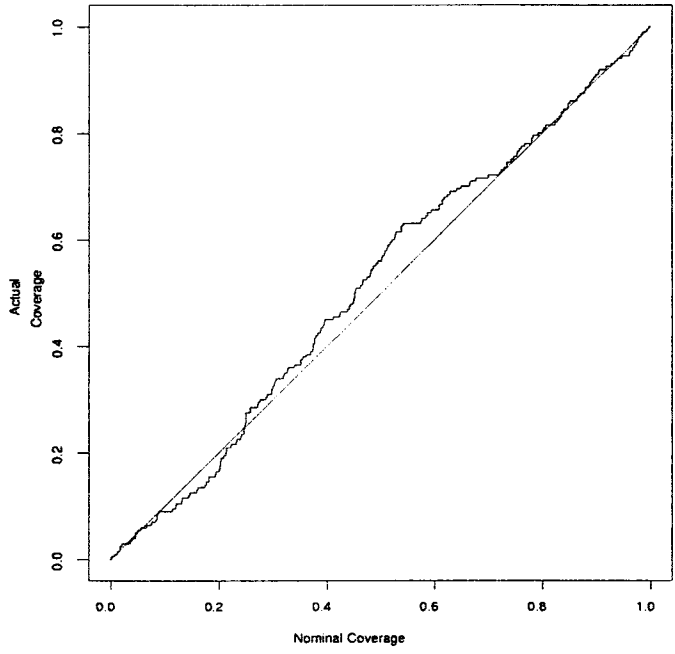

Figure 4 Plot of coverage of $100\alpha\%$ confidence regions as a function of α, for the S-U algorithm using importance sampling, based on results from 200 analyses of the semi-Markov model data.

the conditional distribution of the missing data given the observed data; those problems where it is possible to sample directly from this distribution result in simplified calculation of the score vector and Jacobian matrix. In either case, all quantities required by the S-U algorithm can be calculated recursively, allowing for efficient use of computational resources. Finally, the S-U algorithm can be used when replicates are generated using importance sampling, a case in which existing stochastic approximation methods cannot be used.

Many problems in which the S-U algorithm can be used can also be solved using Monte-Carlo maximum likelihood (Geyer, 1995). Monte-Carlo maximum likelihood has a number of appealing features. Derivative-free methods can be used for non-smooth problems, or for cases when derivative-based methods fail because of poor initial values. Numerical stability far from the MLE may also be enhanced by repeated

use of the same imputed data. However, the $\sqrt{j}$ convergence rate implied by (13) and (27) results in each successive decimal of accuracy in the estimate requiring an expenditure of 100 times the resources required to estimate the previous decimal. Hence, in a sequential calculation such as the S-U algorithm, most of the computation occurs with the current parameter estimates relatively close to the true MLE; a one-sample method such as MCML is unable to make efficient use of replicate data near the true MLE. In addition, sequential methods for Monte-Carlo calculations have intrinsic appeal. The S-U algorithm can be run until the Monte-Carlo error is smaller than a specified size, or until a certain time has elapsed; neither of these stopping rules is possible using MCML. Values of the current parameter estimates can be printed every kth S-U step to give investigators preliminary estimates, as well as to monitor convergence of the algorithm.

In some cases, an entire parametric family of importance distributions may be feasible or appealing. The importance distribution is allowed to change at each S-U step (as long as a limiting distribution is reached); as a result, one may choose parameters of the importance distribution to optimize a criterion based on the variance of the Monte-Carlo error while simultaneously estimating the parameters θ, by adding appropriate estimating equations for the parameters in the importance sampling distribution.

References

Andrews, D. W. K. (1992) Generic uniform convergence. *Econometric Theory* 8, 241-257.

Celeux, G. and Diebold, D. (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Stat. Quarterly*, 2, 73-82.

Crouch, E. A. C., and Spiegelman, D. (1990) The Evaluation of Integrals of the Form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)\,dt$: Application to Logistic-Normal Models, *J. Amer. Statist. Assoc.*, 85, 464-469.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Statist. Soc.* B, 39, 1-37.

Diebold, J. and Ip, E. H. S. (1996) Stochastic EM: method and application. In *Markov Chain Monte Carlo in Practice* (ed.s W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), pp. 259-273. London: Chapman & Hall.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994) *Analysis of Longitudinal Data*, Oxford, UK: Oxford University Press.

Gelfand, A. E. and Carlin, B. P. (1993) Maximum likelihood estimation for constrained or missing data models, *Canadian J. Statist.* **21**, 303-311.

Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23 Symposium on the Interface* (ed. E. M. Keramidas), pp. 156-163. Fairfax: Interface Foundation.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B, **54**, 657-699.

Geyer, C. J. (1995) Estimation and Optimization of Functions. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter), pp. 241-258. London: Chapman & Hall.

Hyde, C. C., and Hall, P. (1980) *Martingale Limit Theory and its Applications*, New York, NY: Academic Press.

Jones, B., and Kenward, M. G. (1989) *Design and Analysis of Cross-Over Studies*, London: Chapman & Hall.

Press, W. H., Teukolsky, S. A., Vetterling, B. T., and Flannery, B. P. (1992) *Numerical Recipes in Fortran: The Art of Scientific Computing (2nd Edition)*, Cambridge, UK: Cambridge University Press.

Ruppert, D., Reish, R. L., Deriso, R. B., and Carroll, R. J. (1984) Optimization using Stochastic Approximation and Monte Carlo Simulation (with Application to Harvesting of Atlantic Menhaden). *Biometrics*, **40**, 535-545.

Ruppert, D. (1991), Stochastic Approximation. In *Handbook of Sequential Analysis* (ed.s B. K. Ghosh and P. K. Sen), pp. 503-29. New York: Marcel Dekker.

Satten, G. A. (1996) Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, **83**, 355-370.

Satten, G. A., Datta, S., and Williamson, J., (1998) Inference based on imputed failure times for the proportional hazards model with interval censored data. *J. Amer. Statist. Assoc.* **93**, 318-327.

Sternberg, M. R. and Satten, G. A. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data. *Biometrics* **55**, 514-522.

Tanner, M. A. (1993) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York: Springer-Verlag.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85**, 699-704.

Wu, C. F. J. (1985) Efficient sequential designs with binary data. *J. Amer. Statist. Assoc.* **80**, 974-984.

Wu, C. F. J. (1986) Maximum likelihood recursion and stochastic approximation in sequential designs. In *Adaptive Statistical Procedures and Related Topics* (ed. J. Van Ryzin), pp. 298-313. Hayward, CA: Institute of Mathematical Statistics.

Younes, L. (1988) Estimation and annealing for gibbsian fields. *Ann. de l'Inst. Henri Poincare. Sect. B, Prob. et Statist.*, **24**, 269-294.

Zeger, S. L., and Karim, M. R., (1991) Generalized linear models with random effects; A Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

Appendix

Regularity assumptions for Theorem 1.

Let $\Omega = [\hat{\theta} - \eta,\ \hat{\theta} + \eta]$, $\eta > 0$ be a compact neighborhood of $\hat{\theta}$, let $\log^+(x) = \max[0, \log(x)]$, $x > 0$, and let $\| \ \|$ denote the Euclidian norm, and let $G_\theta^{-1i}(u \mid y_i) = \inf_x \{G_\theta^i(x \mid y_i) \geq u\}$. Then, the following conditions are assumed to hold for each $1 \leq i \leq N$.

C1. There exist non-negative functions A_i and h_i with $\int_0^1 A_i(u) \log^+[A_i(u)]\, du < \infty$, and $h_i(x) \downarrow 0$ as $x \downarrow 0$, such that

$$\left\| w_\theta^i[G_\theta^{-1i}(u \mid y_i), y_i] - w_{\theta'}^i[G_{\theta'}^{-1i}(u \mid y_i), y_i] \right\|$$

$$\leq A_i(u)\, h_i(\, \|\theta - \theta'\|\,)$$

$\forall\, \theta, \theta' \in \Omega$, $u \in (0,1)$, and

$$E_{G_\theta(x \mid y_i)}[\, w_\theta(X, y_i) \log^+ w_\theta(X, y_i)] < \infty$$

C2a. There exist non-negative functions B_i and k_i with $\int_0^1 B_i(u) \log^+[B_i(u)]\, du < \infty$, and $k_i(x) \downarrow 0$ as $x \downarrow 0$, such that

$$\left\| S_i[G_\theta^{-1i}(u \mid y_i), y_i]\, w_\theta^i[G_\theta^{-1i}(u \mid y_i), y_i] \right.$$
$$\left. - S_i[G_{\theta'}^{-1i}(u \mid y_i), y_i]\, w_{\theta'}^i[G_{\theta'}^{-1i}(u \mid y_i), y_i] \right\| \le B_i(u)\, k_i(\|\theta - \theta'\|)$$

$\forall\, \theta, \theta' \in \Omega$, $u \in (0,1)$, and

$$E_{G_\theta^i(x \mid y_i)}\left[\left\| S_i(X, y_i) \right\| w_\theta^i(X, y_i) \log^+\left\{ \left\| S_i(X, y_i) \right\| w_\theta^i(X, y_i) \right\}\right] < \infty$$

C2b. There exist non-negative functions $\tilde{B}_i$ and $\tilde{k}_i$ with $\int_0^1 \tilde{B}_i(u) \log^+[\tilde{B}_i(u)]\, du < \infty$, and $\tilde{k}_i(x) \downarrow 0$ as $x \downarrow 0$, such that

$$\left\| \tilde{S}_i[G_\theta^{-1i}(u \mid y_i), y_i]\, w_\theta^i[G_\theta^{-1i}(u \mid y_i), y_i] \right.$$
$$\left. - \tilde{S}_i[G_{\theta'}^{-1i}(u \mid y_i), y_i]\, w_{\theta'}^i[G_{\theta'}^{-1i}(u \mid y_i), y_i] \right\| \le \tilde{B}_i(u)\, \tilde{k}_i(\|\theta - \theta'\|)$$

$\forall\, \theta, \theta' \in \Omega$, $u \in (0,1)$, and

$$E_{G_\theta^i(x \mid y_i)}\left[\left\| \tilde{S}_i(X, y_i) \right\| w_\theta^i(X, y_i) \log^+\left\{ \left\| \tilde{S}_i(X, y_i) \right\| w_\theta^i(X, y_i) \right\}\right] < \infty$$

C3. There exist non-negative functions C_i and l_i with $\int_0^1 C_i(u) \log^+[C_i(u)]\, du < \infty$, and $l_i(x) \downarrow 0$ as $x \downarrow 0$, such that

$$\left\| \mathcal{H}_i[G_\theta^{-1i}(u \mid y_i), y_i]\, w_\theta^i[G_\theta^{-1i}(u \mid y_i), y_i] \right.$$
$$\left. - \mathcal{H}_i[G_{\theta'}^{-1i}(u \mid y_i), y_i]\, w_{\theta'}^i[G_{\theta'}^{-1i}(u \mid y_i), y_i] \right\| \le C_i(u)\, l_i(\|\theta - \theta'\|)$$

$\forall\, \theta, \theta' \in \Omega$, $u \in (0,1)$, and

$$E_{G_\theta^i(x \mid y_i)}\left[\left\| \mathcal{H}_i(X, y_i) \right\| w_\theta^i(X, y_i) \log^+\left\{ \left\| \mathcal{H}_i(X, y_i) \right\| w_\theta^i(X, y_i) \right\}\right] < \infty$$

C4. $f_\theta^i(y_i)$, $\mathbb{S}_i(y_i \mid \theta)$ and $\mathbb{H}_i(y_i \mid \theta)$ are twice continuously differentiable in θ, $f_\theta^i(y_i)$ is positive, and $\mathbb{H}_T(\{y_i\} \mid \theta)$ is non-singular on Ω.

Conditions (C1) – (C3) are needed for the uniform SLLN corresponding to the summands w_{ijk}, $S_{ijk}w_{ijk}$, $\tilde{S}_{ijk}w_{ijk}$ and $\mathcal{H}_{ijk}w_{jk}$ respectively. If $G^i_\theta(\cdot \mid y)$ does not depend on θ, these conditions take a simpler form. Condition (C2b) is only required if $S_i(x, y \mid \theta) \neq \tilde{S}_i(x, y \mid \theta)$.

Proof of Theorem 1. For simplicity of presentation, we assume that θ is a real parameter and $N = 1$. In that case, can omit the subscript i throughout and denote $\mathbb{S}_T = \mathbb{S}_1$ by $\mathbb{S}$ etc., throughout the proof. K will stand for a generic constant.

Note that we can write $w_{ijk} = w^i_{\theta_j}[G^{-1i}_{\theta_j}(U_{jk}, y), y]$, where U_{ijk} are i.i.d. $U[0,1]$. It is easy to see that by (C1), the conditions BD, P-SLLN and S-LIP in the uniform SLLN of Andrews (1992) are satisfied; here, uniformity refers to uniform in $\theta \in \Omega$. Hence, by Theorem 3 of the same paper,

$$\frac{1}{M}\sum_{k=1}^{M} w_{ijk} \xrightarrow{a.s.} \phi^i_{\theta_j}(y) \ ,$$

uniformly in $j \geq 1$, as $M \to \infty$, provided the θ_j's lie in Ω; $\phi^i_\theta(\cdot)$ is defined in (12). In the same way, by (C2) and (C3),

$$\frac{1}{M}\sum_{k=1}^{M} S_{jk}\, w_{jk} \xrightarrow{a.s.} \phi_{\theta_j}(y)\, \mathbb{S}(y \mid \theta_j) \ ,$$

$$\frac{1}{M}\sum_{k=1}^{M} \tilde{S}_{jk}\, w_{jk} \xrightarrow{a.s.} \phi_{\theta_j}(y)\, \tilde{\mathbb{S}}(y \mid \theta_j) \ ,$$

and

$$\frac{1}{M}\sum_{k=1}^{M} \mathcal{H}_{jk}\, w_{jk} \xrightarrow{a.s.} \phi_{\theta_j}(y)\Big[\, \mathbb{H}(y \mid \theta_j) \ - \ \mathbb{S}(y \mid \theta_j)\, \tilde{\mathbb{S}}^{\tau}(y \mid \theta_j)\Big] \ ,$$

uniformly in $j \geq 1$, as $M \to \infty$, provided the θ_j's lie in Ω. Therefore

$$r_j \equiv \left| \frac{\hat{\mathbb{S}}_j}{\hat{\mathbb{H}}_j} \ - \ \frac{j^{-1}\sum_{j'} \phi_{\theta_{j'}}(y)\, \mathbb{S}(y \mid \theta_{j'})}{j^{-1}\sum_{j'} \phi_{\theta_{j'}}(y)\, \mathbb{H}(y \mid \theta_{j'})} \right| \xrightarrow{a.s.} 0, \qquad \text{(A1.1)}$$

uniformly in $j \geq 1$, as $M \to \infty$, provided the θ_j's lie in Ω.

Next, note that by Taylor expanding both the numerator and the denominator of the second ratio in r_j we get

$$\left| \frac{j^{-1}\sum_{j'=1}^{j} \phi_{\theta_{j'}}(y)\, \mathbb{S}(y \mid \theta_{j'})}{j^{-1}\sum_{j'=1}^{j} \phi_{\theta_{j'}}(y)\, \mathbb{H}(y \mid \theta_{j'})} - \frac{\mathbb{S}(y \mid \bar{\theta})}{\mathbb{H}(y \mid \bar{\theta})} \right| \leq K\, j^{-1} \sum_{j'=1}^{j} (\theta_{j'} - \bar{\theta})^2 . \tag{A1.2}$$

Similarly,

$$\left| j^{-1}\sum_{j'=1}^{j} \frac{\mathbb{S}(y \mid \theta_{j'})}{\mathbb{H}(y \mid \theta_{j'})} - \frac{\mathbb{S}(y \mid \bar{\theta})}{\mathbb{H}(y \mid \bar{\theta})} \right| \leq K\, j^{-1} \sum_{j'=1}^{j} (\theta_{j'} - \bar{\theta})^2 . \tag{A1.3}$$

Combining (A1.1), (A1.2) and (A1.3) we get

$$\begin{aligned} \left| \frac{\widehat{\mathbb{S}}_j}{\widehat{\mathbb{H}}_j} - j^{-1}\sum_{j'=1}^{j} \frac{\mathbb{S}(y \mid \theta_{j'})}{\mathbb{H}(y \mid \theta_{j'})} \right| &\leq r_j + K\, j^{-1} \sum_{j'=1}^{j} (\theta_{j'} - \bar{\theta})^2 \\ &\leq r_j + K\, j^{-1} \sum_{j'=1}^{j} (\theta_{j'} - \widehat{\theta})^2 , \end{aligned} \tag{A1.4}$$

provided the θ_j's lie in Ω.

By a Taylor expansion of $\mathbb{S}(y \mid \widehat{\theta})$ around θ_j, we get

$$0 = \mathbb{S}(y \mid \widehat{\theta}) = \mathbb{S}(y \mid \theta_j) + (\widehat{\theta} - \theta_j)\, \mathbb{H}(y \mid \theta_j) + \tfrac{1}{2}\, (\widehat{\theta} - \theta_j)^2\, \mathbb{S}''(y \mid \theta_j^*) , \tag{A1.5}$$

where θ_j^* lies between θ_j and $\widehat{\theta}$. Dividing (A1.5) by $\mathbb{H}(y \mid \theta_j)$ and averaging, we obtain

$$\widehat{\theta} = \bar{\theta}_j - j^{-1} \sum_{j'=1}^{j} \frac{\mathbb{S}(y \mid \theta_{j'})}{\mathbb{H}(y \mid \theta_{j'})} - (2j)^{-1} \sum_{j'=1}^{j} (\widehat{\theta} - \theta_{j'}) \frac{\mathbb{S}''(y \mid \theta_{j'}^*)}{\mathbb{H}(y \mid \theta_{j'})} . \tag{A1.6}$$

Using $\theta_{j+1} = \bar{\theta}_j - \widehat{\mathbb{S}}_j / \widehat{\mathbb{H}}_j$, (A1.4) and (A1.6) we obtain

$$| \widehat{\theta} - \theta_{j+1} | \leq r_j + K\, j^{-1} \sum_{j'=1}^{j} (\theta_{j'} - \widehat{\theta})^2 ,$$

provided the θ_j's lie in Ω.

Let $P > \max(2K, \eta^{-1})$. Then re-write the above inequality as

$$a_{j+1} \leq \epsilon_j + \frac{K}{P} j^{-1} \sum_{j'=1}^{j} a_{j'}^2 \quad , \tag{A1.7}$$

where $a_j = P\,|\hat{\theta} - \theta_j|$ and $\epsilon_j = r_j P$, $j \geq 1$. By (A1.1), on a set of probability one, select M large enough so that $\epsilon_j < \frac{1}{2}$, $\forall j \geq 1$, provided the θ_j's lie in Ω. Suppose that the starting value $\theta_1 \in \Omega = [\hat{\theta} - \eta, \hat{\theta} + \eta]$. Then $a_1 < 1$ and using (A1.7) we find that $a_2 \leq (\frac{1}{2} + \frac{K}{P}) < 1$ which implies that $\theta_2 \in \Omega$. Inductively using (A1.7) we find that $a_j \in \Omega$ and $0 \leq a_j \leq (\frac{1}{2} + \frac{K}{P})$, $\forall j \geq 1$. Therefore, applying "limsup" to both sides of (A1.7) and using the fact that for $a_j \geq 0$,

$$\limsup_{j \to \infty} j^{-1} \sum_{j'=1}^{j} a_{j'}^2 \leq (\limsup_{j \to \infty} a_j)^2 \quad ,$$

and that $r_j \overset{a.s.}{\to} 0$ as $j \to \infty$ (for any given replication size M), we conclude that on a set of probability one,

$$\limsup_{j \to \infty} a_j \leq \frac{1}{2} (\limsup_{j \to \infty} a_j)^2 \quad ,$$

which immediately implies $\limsup_{j \to \infty} a_j = 0$, since $\limsup_{j \to \infty} a_j \in [0,1)$. Hence, convergence of θ_j to $\hat{\theta}$ is established. For the case $N > 1$ note that while equations (A1.1) - (A1.3) become considerably more complicated, equation (A1.4) remains unchanged. □

Regularity assumptions for Theorem 2: Assume the conditions for Theorem 1 plus

N1. The matrix $\sum_i V^i$ is positive definite.

Proof of Theorem 2. Write

$$\sqrt{j}\,(\theta_{j+1} - \hat{\theta}) = \sqrt{j}\,(\bar{\theta}_j - \hat{\theta} - \hat{\mathbb{H}}_j^{-1} \hat{\mathbb{S}}_j)$$

$$= \hat{\mathbb{H}}_j^{-1} \{ \mathbb{H}_T(\{y_i\} \mid \hat{\theta}) - \hat{\mathbb{H}}_j \} \sqrt{j} \, (\bar{\theta}_j - \hat{\theta})$$

$$+ \hat{\mathbb{H}}_j^{-1} [\sqrt{j} \{ \mathbb{H}_T(\{y_i\} \mid \hat{\theta})(\bar{\theta}_j - \hat{\theta}) - j^{-1} \sum_{j'=1}^{j} \mathbb{S}_T(\{y_i\} \mid \theta_{j'})\}]$$

$$- \hat{\mathbb{H}}_j^{-1} \sqrt{j} \{ \hat{\mathbb{S}}_j - j^{-1} \sum_{j'=1}^{j} \mathbb{S}_T(\{y_i\} \mid \theta_{j'})\} \ . \qquad (A2.1)$$

Since $\hat{\mathbb{H}}_j \xrightarrow{P} \mathbb{H}_T(\{y_i\} \mid \hat{\theta})$ and $\sqrt{j} \, (\bar{\theta}_j - \hat{\theta})$ is $O_p(1)$, the first term of (A2.1) converges to zero in probability.

Taylor expanding $\mathbb{S}_T(\{y_i\} \mid \theta_j)$ around $\hat{\theta}$ and averaging, we obtain

$$j^{-1} \sum_{j'=1}^{j} \mathbb{S}_T(\{y_i\} \mid \theta_{j'}) = \mathbb{H}_T(\{y_i\} \mid \hat{\theta}) \, (\bar{\theta}_j - \hat{\theta}) + O_p(j^{-1} \sum_{j'=1}^{j} | \theta_{j'} - \hat{\theta} |^2) \ ;$$

therefore, the second term on the right hand side of (A2.1) is $o_p(1)$, since $j^{1/2} \sum_{j'=1}^{j} | \theta_{j'} - \hat{\theta} | = O_p(1)$.

From (5), (6) and (8), we have

$$\hat{\mathbb{S}}_j \equiv \sum_{i=1}^{N} \frac{\frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} S_{ij'k} \, w_{ij'k}}{\frac{1}{j \cdot M} \sum_{j'=1}^{j} \sum_{k=1}^{M} w_{ij'k}} \equiv \sum_{i=1}^{N} \frac{N_i}{D_i} \ .$$

It is easy to verify the conditional Lindeberg condition for linear combinations of (N_i, D_i). By the martingale central limit theorem (Corollary 3.1 of Hyde and Hall, 1980), the asymptotic distribution of $\sqrt{j} \sum_{i=1}^{N} (N_i, D_i)$ is the multivariate normal distribution with mean vector

$$\sqrt{j} \left(\sum_{i=1}^{N} j^{-1} \sum_{j'=1}^{j} \mathbb{S}_i(y_i \mid \theta_{j'}) \, \phi^i_{\theta_{j'}}(y_i), \ \sum_{i=1}^{N} j^{-1} \sum_{j'=1}^{j} \phi^i_{\theta_{j'}}(y_i) \right)$$

and variance-covariance matrix $M^{-1} \sum_{i=1}^{N} V^i$, where the V^i is defined in (15)-(18). By the delta method, we have

$$\sqrt{j} \, (\hat{\mathbb{S}}_j - \mu_j) \xrightarrow{d} N(0, \mathbb{V}) \ , \qquad (A2.2)$$

where
$$\mu_j = \sum_{i=1}^{N} \frac{j^{-1}\sum_{j'=1}^{j} \mathbb{S}_i(y_i \mid \theta_{j'})\, \phi^i_{\theta_{j'}}(y_i)}{j^{-1}\sum_{j'=1}^{j} \phi^i_{\theta_{j'}}(y_i)}$$

and where $\mathbb{V}$ is given in equation (19). Using a Taylor series argument as in the proof of Theorem 1, we get for each i

$$j^{-1}\sum_{j'=1}^{j} \frac{\mathbb{S}_i(y_i \mid \theta_{j'})\, \phi^i_{\theta_{j'}}(y_i)}{j^{-1}\sum_{j'=1}^{j} \phi^i_{\theta_{j'}}(y_i)} = j^{-1}\sum_{j'=1}^{j} \mathbb{S}_i(y_i \mid \theta_{j'}) + O_p(j^{-1}\sum_{j'=1}^{j} (\theta_{j'} - \widehat{\theta})^2) \ ;$$

hence,

$$\sqrt{j}\left\{ \mu_j - j^{-1}\sum_{j'=1}^{j} \mathbb{S}(\{y_i\} \mid \theta_{j'}) \right\} = o_p(1) \ . \qquad \text{(A2.3)}$$

Combining (A2.1), (A2.2) and (A2.3) we find $\sqrt{j}(\theta_{j+1} - \widehat{\theta}) \xrightarrow{d} N(0, \Sigma)$, where

$$\Sigma = \mathbb{H}^{-1}(\{y_i\} \mid \widehat{\theta}) \cdot \mathbb{V} \cdot \mathbb{H}^{-T}(\{y_i\} \mid \widehat{\theta}) \ . \quad \square$$

Reproduced with permission of copyright owner.
Further reproduction prohibited without permission.