A Stochastic Approximation Algorithm with Markov Chain Monte-Carlo Method for Incomplete Data Estimation Problems

Ming Gao Gu[1] and Fan Hui Kong[2]

April, 1998

ABSTRACT

We propose a general procedure for solving incomplete data estimation problems. The procedure can be used to find the maximum likelihood estimate (MLE) or to solve estimating equations in difficult cases such as estimation with the censored or truncated regression model, the nonlinear structural measurement error model and the random effects model. The procedure is based on the general principle of stochastic approximation and the Markov chain Monte-Carlo method. Applying the theory on adaptive algorithms, we derive conditions under which the proposed procedure converges. Simulation studies also indicate that the proposed procedure consistently converges to the MLE for the structural measurement error logistic regression model.

AMS 1980 subject classification. Primary 62E20; secondary 62G05, 60G15.

Other Keywords: Incomplete data, Markov chain Monte-Carlo, Maximum likelihood estimation, Measurement error, Stochastic approximation.

[1]Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada H3A 2K6

[2]Westat, 1441 West Montgomery Avenue, Rockville, MD 20850, U.S.A.

1 Introduction.

Suppose we have a random vector $X \in E$, where E is some sample space. Further assume that X has a density function $f(x|\theta)$, which depends on a d-dimensional parameter $\theta \in \Theta$. Many statistical estimation procedures, such as the maximum likelihood estimation, estimating equation and robust regression procedures, share the common feature of finding a $\hat{\theta} \in \Theta$ such that

$$H(\hat{\theta}, X) = 0, \tag{1}$$

for a given function $H(\cdot, \cdot) : \Theta \times E \to R^d$. In the case of maximum likelihood estimation, $H(x, \theta) = \nabla_\theta \log f(x|\theta)$, where ∇ denotes the gradient operator with respect to θ. This estimating procedure is widely used in practice and its properties have been well studied.

In practice, however, data X may not be completely observable; instead only a proxy, Y, is observable. For example, in survival analysis, some data may be subject to right censoring. Other types of incomplete data may not be categorized so straightforwardly. For example, in error-in-variable and mixed model problems, the model contains a latent variable that cannot be observed.

When data are incomplete, the following procedure is usually employed. Suppose that only the incomplete data $Y \in E'$ is observed. Also suppose that given $Y = y$, the complete data X, which cannot be determined in certainty, follows the conditional distribution with density $\pi_\theta(x|y)$, depending on θ. To estimate the parameter θ, one usually finds $\hat{\theta} \in \Theta$ such that

$$h(\hat{\theta}) = h(\hat{\theta}, Y) = 0, \tag{2}$$

where function h is defined by

$$h(\theta, y) = E_\theta \left[H(\theta, X) | Y = y \right] = \int H(\theta, x) \pi_\theta(x|y) dx. \tag{3}$$

In a recent paper [1] dealing with estimation in a censored regression model, Lai and Ying have called the procedure given by (2) and (3) the missing information principle.

The major obstacle in implementing this procedure is that the function h given by (3) usually has no closed analytic expression, and therefore finding $\hat{\theta}$ in (2) is difficult. One way to overcome this obstacle is to use numerical integration in (3). In attempting to find the MLE of error-in-variable logistic regression model, Schafer [2] found that this approach gives unstable estimates and is usually inferior to other approaches to the problem.

Another way to overcome this obstacle is to use Monte-Carlo integration to approximate $h(\theta)$ ([3], [4]). The advantage of this approach is that the approximation is not affected by the "curse of dimensionality." At the same time the Metropolis-Hastings algorithm ([5], [6]) can be used to simulate data from the conditional distribution $\pi_\theta(x|y)$. In many cases, the conditional distribution $\pi_\theta(x|y)$ admits the form $C(y,\theta)g(x|y,\theta)$, where g has an analytic form but $C^{-1}(y,\theta) = \int g(x|y,\theta)dy$, the normalizing constant, has no closed analytic expression. The Metropolis-Hastings algorithm is well suited for such problems. Examples that bear this characteristic can be found in [7], where rank based regression procedure is proposed for interval censored data, and in [4], where procedures for deriving MLE in a nonlinear mixed model are discussed. Another example is maximum likelihood estimation with the error-in-variable logistic regression model as discussed in detail in Section 4.

The question still remains as to how close one should approximate the function $h(\theta)$. In order to derive an accurate value of $\hat{\theta}$, one needs a large number of simulations, especially when θ is in the neighborhood of $\hat{\theta}$. Stochastic approximation, first proposed by Robbins and Monro in [8], provides an answer to this question. By recent developments on adaptive control and stochastic approximation ([9]), the Metropolis-Hastings algorithm can be incorporated into the simulation step to form a general procedure for this task.

The aim of this paper is to propose a procedure for finding $\hat{\theta}$ in the general statistical model of (1), (2) and (3). We show that the proposed procedure converges under mild conditions. We also illustrate our procedure using the example of maximum likelihood estimation in the error-in-variable logistic regression model. Simulation results indicate that the proposed procedure works well for this model.

2 The Proposed Algorithm.

In order to introduce our algorithm, we first introduce the Markov transition probability $\Pi_\theta(x, A)$. To use the Metropolis-Hastings algorithm ([5], [6]) to simulate values from the conditional distribution $\pi_\theta(x|y)$, we can construct a Markov transition probability $\Pi_\theta(x, A)$ such that $\pi_\theta(x|y)$ is the unique invariant distribution on E. In other words, for any measurable set A in E,

$$\int_A \pi_\theta(dx|y) = \int_E \pi_\theta(dx|y)\Pi_\theta(x, A).$$

In the case that $\pi_\theta(x|y) = C(y, \theta)g(x|y, \theta)$, where g has a closed analytic expression, [6] suggested the use of

$$\Pi_\theta(x, dx') = q(x, x') \min\left\{\frac{g(x'|y, \theta)\ q(x', x)}{g(x|y, \theta)\ q(x, x')}, 1\right\} dx'$$

for $x' \neq x$ and $\Pi_\theta(x, \{x\}) = 1 - \int_{z\neq x} \Pi_\theta(x, dz)$, where $q(x, x')$ is any aperiodic recurrent transition density. In our case, $q(x, x')$ can be chosen to depend on θ and y.

To apply the stochastic approximation algorithm more efficiently, we need to approximate the derivative matrix $-(\partial/\partial\theta)h(\theta)$. In the algorithm proposed below, we assume that there exists a matrix function $I(\theta, x)$ such that

$$G(\theta) \equiv \mathrm{E}_\theta\left[I(\theta, X)|Y = y\right] \tag{4}$$

is close to $-(\partial/\partial\theta)h(\theta)$ for θ in the neighborhood of $\hat{\theta}$. In Section 3 we will discuss how to choose such a function $I(\theta, x)$.

Following the general principle of stochastic approximation, we choose a positive integer m and a sequence of positive constants $\{\gamma_k\}$ such that

(C.1) $\sum_{k=1}^{\infty} \gamma_k = \infty$ and

(C.2) $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.

We propose the following algorithm for finding the $\hat{\theta}$ that satisfies (2).

A stochastic approximation algorithm for estimation with the Markov chain Monte-Carlo method

Step 0. Choose initial values $\theta_0 \in \Theta$, Γ_0 and $X_{0,m} \in E$ and set $k = 1$.

Step 1. For fixed k, set $X_{k,0} = X_{k-1,m}$. For $i = 1, \cdots, m$, simulate $X_{k,i}$ from the transition probability $\Pi_{\theta_{k-1}}(X_{k,i-1}, \cdot)$.

Step 2. Update the estimate of $\hat{\theta}$ by

$$
\begin{aligned}
\Gamma_k &= \Gamma_{k-1} + \gamma_k \left(\bar{I}(\theta_{k-1}, \boldsymbol{X}_k) - \Gamma_{k-1} \right) \quad \text{and} \\
\theta_k &= \theta_{k-1} + \gamma_k \Gamma_k^{-1} \bar{H}(\theta_{k-1}, \boldsymbol{X}_k),
\end{aligned} \tag{5}
$$

where $\boldsymbol{X}_k = (X_{k,1}, \cdots, X_{k,m})$ and

$$
\begin{aligned}
\bar{H}(\theta, \boldsymbol{x}) &= \frac{1}{m} \left\{ H(\theta, X_1) + \ldots + H(\theta, X_k) \right\}, \quad (6) \\
\bar{I}(\theta, \boldsymbol{x}) &= \frac{1}{m} \left\{ I(\theta, X_1) + \ldots + I(\theta, X_m) \right\}.
\end{aligned}
$$

Set $k = k + 1$ and go to Step 1 until the sequence $\{\theta_k\}$ converges.

The choice for the sequence $\{\gamma_k\}$ is usually $\{1/k\}$.

The choice of m should not affect the convergence of the proposed procedure. However, a good choice of m may make the procedure more stable and save computation time. In general, when the dimension of the parameter space Θ is high, it may be computationally expensive to invert the matrix Γ_k in (5). In this case, one may choose a larger m. A larger m makes $\bar{H}(\theta_{k-1}, \boldsymbol{X}_k)$ smoother and closer to $h(\theta_{k-1})$ and $\bar{I}(\theta_{k-1}, \boldsymbol{X}_k)$ closer to $G(\theta_{k-1})$ and therefore shortens the total number of iterations for $\{\theta_k\}$ to converge. Our experience shows that m can be in the range of 10 to 100.

3 Choices of Function $I(\theta, x)$.

A good choice of the function $I(\theta, x)$ will increase the rate of convergence of the stochastic approximation algorithm. To obtain such a choice, we need to generalize the missing information theorem of [10]. Note that we assume that X comes from the family $\{f(x|\theta), \theta \in \Theta\}$.

Lemma 1 *Suppose that $H(\theta, x)$ is twice differentiable with respect to θ and that the order of integration and differentiation can be exchanged. Then for $h(\theta)$ defined in (3), we have*

$$-\frac{\partial}{\partial\theta}h(\theta) = -E_\theta\left[\frac{\partial}{\partial\theta}H(\theta, X)\Big|Y = y\right] - Cov_\theta\left(H(\theta, X), \frac{\partial}{\partial\theta}\log f_\theta(X)\Big|Y = y\right). \qquad (7)$$

In the case of maximum likelihood estimation, $H(\theta, x) = (\partial/\partial\theta)\log f_\theta(x)$, Lemma 1 reduces to Louis' missing information theorem ([10]). Lemma 1 can be proved with the same argument as in [10]. See also [11], page 75. The proof is omitted here.

According to [9], a good choice of function $I(\theta, x)$ would be a function such that $\mathrm{E}[I(\theta, X)|Y = y] = -(\partial/\partial\theta)h(\theta)$. According to Lemma 1, we can choose

$$I(\theta, x) = -\frac{\partial}{\partial\theta}H(\theta, x) - H(\theta, X)\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^t, \qquad (8)$$

where a^t denotes the transpose of vector a. In this case,

$$\begin{aligned} G(\theta) &= -\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}H(\theta, X)\Big|Y = y\right] - \mathrm{E}_\theta\left[H(\theta, X)\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^t\Big|Y = y\right] \\ &= -\frac{\partial}{\partial\theta}h(\theta) - h(\theta)\mathrm{E}_\theta^t\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\Big|Y = y\right]. \end{aligned}$$

In the neighborhood of $\hat{\theta}$, defined by (2), $h(\theta)$ is small and therefore $G(\theta)$ is close to $-(\partial/\partial\theta)h(\theta)$. Moreover, at $\theta = \hat{\theta}$, $G(\hat{\theta}) = -(\partial/\partial\theta)h(\hat{\theta})$. In Section 5, we will see that under general conditions, $\Gamma_k \to G(\hat{\theta})$ as $\theta_k \to \hat{\theta}$. Thus when k is large, Γ_k can serve as an estimate of $-(\partial/\partial\theta)h(\hat{\theta})$. In the case of $H(\theta, x) = (\partial/\partial\theta)\log f_\theta(x)$, Γ_k^{-1} is an estimate of the covariance matrix of $\hat{\theta}$.

4 Application to the Measurement Error Model.

In this section, we apply the stochastic approximation algorithm to find the MLE in the logistic regression model, with covariates measured with errors. Measurement error problems often arise in epidemiologic studies when risk factors cannot be measured accurately. A detailed introduction to this subject can be found in [12] or [13]. Different estimation methods were proposed for this problem. Schafer ([2]) was the first tried to calculate the exact MLE with an approximate EM algorithm.

Suppose that the true, unobservable covariates $Z_1, \ldots, Z_n$ are i.i.d. from a population with a known density function $f_Z(z)$. The observed data are $(U_i, \boldsymbol{V}_i)$, $i = 1, \ldots, n$. Given that $Z_i = z_i$, U_i is binary with $\Pr\{U_i = 0 | Z_i = z_i\} = 1 - \Pr\{U_i = 1 | Z_i = z_i\}$, and

$$\Pr\{U_i = 1 | Z_i = z_i\} = \{1 + \exp(\alpha + \beta z_i)\}^{-1}, \tag{9}$$

and V_i is distributed as $N(z_i, \sigma^2)$. The log-likelihood function for parameter (α, β) is

$$l(\alpha, \beta) = \sum_{i=1}^{n} \int \left[(1 - U_i)(\alpha + \beta z_i) - \log\{1 + \exp(\alpha + \beta z_i)\}\right] \pi(z_i | U_i, V_i, \alpha, \beta) dz_i, \tag{10}$$

where $\pi(z|u, v, \alpha, \beta)$ is the conditional density of Z_i given $U_i = u$, $V_i = v$ and parameter (α, β). In this case,

$$\pi(z|u, v, \alpha, \beta) = \frac{1}{C} f_Z(z) \phi(\frac{v - z}{\sigma}) \frac{\exp\{(1 - u)(\alpha + \beta z)\}}{\{1 + \exp(\alpha + \beta z)\}}, \tag{11}$$

where C is the normalizing constant so that the integral of $\pi(z|u, v, \alpha, \beta)$ with respect to z is 1.

To apply the proposed stochastic approximation algorithm to find the MLE, $(\hat{\alpha}, \hat{\beta})$, we use the Metropolis-Hastings algorithm to construct a Markov kernel with its invariant measure $\pi(z|y, x, \alpha, \beta)$ (see the discussion at the beginning of Section 2). One possibility is to take $q(z, z^*)$ to be $f_Z(z^*)\phi((v - z^*)/\sigma)/C'$, where C' is the normalizing constant so that $q(z, z^*)$ is a density in z^*. With this choice, the resulting Metropolis algorithm is closely related to the acceptance/rejection methods. Hastings [6] called this the method of independence chain. The resulting transition probability takes the simple form

$$\Pi_{(\alpha,\beta)}(z, dz^*) = \frac{f_Z(z^*)}{C'} \phi(\frac{v - z^*}{\sigma}) \min\left\{\frac{\mathrm{e}^{(1-u)(\alpha+\beta z^*)}\ (1 + \mathrm{e}^{\alpha+\beta z})}{\mathrm{e}^{(1-u)(\alpha+\beta z)}\ (1 + \mathrm{e}^{\alpha+\beta z^*})}, 1\right\} dz^* \tag{12}$$

for $z^* \neq z$ and $\Pi_{(\alpha,\beta)}(z, \{z\}) = 1 - \int_{w \neq z} \Pi_{(\alpha,\beta)}(z, dw)$. The function H is the score function of the complete data. $H(\alpha, \beta, \boldsymbol{Z}) = \nabla L(\alpha, \beta, \boldsymbol{Z})$, where ∇ denotes the gradient operator and $L(\alpha, \beta, \boldsymbol{Z})$ is the log-likelihood function of the complete data,

$$L(\alpha, \beta, \boldsymbol{Z}) = \sum_{j=1}^{n} \left[(1 - U_j)(\alpha + \beta Z_j) - \log\{1 + \exp(\alpha + \beta Z_j)\}\right].$$

According to the discussion in Section 3, function $I(\alpha, \beta, \boldsymbol{Z})$ takes the form

$$-\nabla^2 L(\alpha, \beta, \boldsymbol{Z}) - \nabla L(\alpha, \beta, \boldsymbol{Z}) \left\{\nabla L(\alpha, \beta, \boldsymbol{Z})\right\}^t ,$$

where ∇^2 is the Hessian operator.

We have carried out a simulation study using the proposed stochastic approximation procedure for the case $\alpha = 0$, $\beta = 1$, $n = 200$. The distribution of Z is $\mathrm{N}(0, 1)$ and $\sigma^2 =$ .10, .25, .50 and .75 respectively. The maximum number of iterations of the stochastic approximation procedure is set at $K = 50$ in each simulation. The number m is set at $m = 50$. We use $\gamma_k = 1/k$. The initial values are set at $\alpha_0 = 0.5$, $\beta_0 = 0.5$ and $\Gamma_0 = 0$. The following table gives the mean square errors of $\hat{\alpha}$ and $\hat{\beta}$ and the average of the corresponding diagonal values of Γ_K^{-1}. The result is based on 2000 simulations. The numbers in Table 1 are multiplied by 1000 for easy presentation.

Table 1. $1000 \times MSE$s and the the corresponding average of the diagonal elements of Γ_K^{-1} based on 2000 simulations

σ^2	MSE($\hat{\alpha}$)	Average of $\Gamma_K^{-1}(1, 1)$	MSE($\hat{\beta}$)	Average of $\Gamma_K^{-1}(2, 2)$
0.10	25	25	42	40
0.25	25	25	48	48
0.50	26	26	59	59
0.75	27	26	63	67

A common feature of the previous estimation methods proposed for the nonlinear error-in-variable models is that they give accurate estimate only when the variance of the measurement error σ^2 is small. When σ^2 becomes large, such as .75, these methods give unstable results ([2]). Table 1 suggests that our procedure give accurate MLEs while the mean square error grows proportionally with σ^2. Note that only in the limiting case (or when sample size n becomes very large) is the mean square error of the MLE equal to the inverse of the information matrix. So the small discrepancies between the second and the third columns and the fourth and the last columns are very reasonable.

In a practical application, the maximum number of iterations, K, may be sequentially determined accoring to the sequence $\{\theta_k, k < K\}$. For convenience, we have set this number at 50 in our simulation study. This does not seems cause any particular problem since at iteration 50, all sequences θ_k have already converged in the parameter setting given above. In fact, we have run the simulations for $K = 70, 100$ (with all other factors the same as in $K = 50$) and the results obtained are practically the same as those given in Table 1.

5 Convergence Theorem.

In this section we formulate a convergence theorem for the proposed algorithm. Theorem 1 is based on Theorem 3.17 (page 304) of [9], which gives the conditions under which the stochastic approximation algorithm converges for observations from Markov chains as described in Section 2. We give conditions that are more transparent and easier to verify. Suppose that Θ is an open set in R^d and E is an open set in R^n.

Conditions on the γ_k:

Suppose that the sequence $\{\gamma_k, K \geq 1\}$ satisfies (C.1) and (C.2) in Section 2.

Conditions on the transition probability Π_θ:

Let Q be any compact subset of Θ and let $q > 1$ be a sufficiently large real number. The constants $C_1, \ldots, C_4$ and λ in the following may depend on Q and q.

(C.3) **Integrability.** There exists a C_1 such that for any $x \in E$, $\theta \in \Theta$ and $k \geq 1$,

$$\int (1 + |y|^q)\, \Pi_\theta^k(x, dy) \leq C_1 \left(1 + |x|^q\right).$$

In the above and in the following, $\Pi_\theta^k(x, dy) = \int \cdots \int \Pi(x, dx_1) \cdots \Pi(x_{k-2}, dx_{k-1})\Pi(x_{k-1}, dy)$ and $|x|$ denotes the length of vector x.

(C.4) **Convergence of the Markov chains.** Let π_θ be the unique invariant measure associated with Π_θ. For every $\theta \in D$,

$$\lim_{k\to\infty} \sup_{x\in E} \frac{1}{1+|x|^q} \int (1+|y|^q) \left|\Pi_\theta^k(x, dy) - \pi_\theta(dy)\right| = 0.$$

(C.5) **Continuity in θ.** There exist constants C_2 and C_3, such that for all $\theta, \theta' \in Q$

$$\begin{aligned}\left|\int (1+|y|^q)\left\{\Pi_\theta(x, dy) - \Pi_{\theta'}(x, dy)\right\}\right| &\leq C_2|\theta - \theta'|\,(1+|x|^q)\,;\\ \left|\int (1+|y|^q)\left\{\pi_\theta(dy) - \pi_{\theta'}(dy)\right\}\right| &\leq C_3|\theta - \theta'|.\end{aligned}$$

(C.6) **Continuity in x.** There exists constant C_4, such that for $\theta, \theta' \in Q$ and $x_1, x_2 \in E$

$$\sup_{\theta\in\Theta}\left|\int (1+|y|^{q+1})\left\{\Pi_\theta(x_1, dy) - \Pi_\theta(x_2, dy)\right\}\right| \leq C_4|x_1 - x_2|\,(1+|x_1|^q + |x_2|^q)\,.$$

Conditions on functions $H(\theta, x)$ and $I(\theta, x)$:

(C.7) For any compact subset Q of Θ, there exist positive constants p, K_1, K_2, K_3 and $\lambda' > 1/2$ such that for all $x, x_1, x_2 \in E$ and $\theta, \theta' \in Q$,

$$\begin{aligned}|H(\theta, x)| &\leq K_1\left(1+|x|^{p+1}\right);\\ |H(\theta, x_1) - H(\theta, x_2)| &\leq K_2|x_1 - x_2|\,(1+|x_1|^p + |x_2|^p)\,;\\ |H(\theta, x) - H(\theta', x)| &\leq K_3|\theta - \theta'|^{\lambda'}\left(1+|x|^{p+1}\right).\end{aligned}$$

The same inequalities hold for $I(\theta, x)$.

Since $Y = y$ is fixed in this investigation, we can suppress the variable y in $\pi_\theta(x|y)$. Therefore, we can write $h(\theta) = \int H(\theta, x)\pi_\theta(dx)$ and $G(\theta) = \int I(\theta, x)\pi_\theta(dx)$. Consider the solution $(\theta(t), \Gamma(t))$, $t \geq 0$ of the ordinary differentiate equation (ODE)

$$\begin{pmatrix} \frac{d}{dt}\theta(t) \\ \frac{d}{dt}\Gamma(t) \end{pmatrix} = \begin{pmatrix} \Gamma(t)^{-1}h(\theta(t)) \\ G(\theta(t)) - \Gamma(t) \end{pmatrix}, \quad \begin{pmatrix} \theta(0) \\ \Gamma(0) \end{pmatrix} = \begin{pmatrix} z \\ \Gamma \end{pmatrix}. \tag{13}$$

A point (z^*, Γ^*) is called a stability point if the ODE (13) admits the only solution $\theta(t) = z^*$, $G(\theta(t)) = \Gamma^*$, $t \geq 0$ if $\theta(0) = z^*$, $\Gamma(0) = \Gamma^*$. It is easy to see that $(\hat{\theta},\, G(\hat{\theta}))$ is a stability point of the ODE (13). A set D is called a domain of attraction of a stability point (z^*, Γ^*) if the solution of (13) with $(\theta(0), \Gamma(0)) \in D$ remains indefinitely in D and converges to (z^*, Γ^*).

Theorem 1 *Assume that the conditions (C.1-C.7) are valid. If $\{(\theta_k, \Gamma_k), k \geq 1\}$, defined by (5), is a bounded sequence and visits infinitely often a compact subset of the domain of attraction of the stability point $(\hat{\theta}, G(\hat{\theta}))$ of ODE (13) almost surely, then*

$$\theta_k \to \hat{\theta} \text{ and } \Gamma_k \to G(\hat{\theta}) \quad \text{almost surely.} \tag{14}$$

Condition (C.3) is a moment condition. Condition (C.4) is a stronger version of the assumption that the Markov chain driven by the transition probability $\Pi_\theta(x, dy)$ converges uniformly to the invariant measure π_θ. Nevertheless, such a condition is usually satisfied whenever $\pi(\theta)$ has high moments. Conditions (C.5) and (C.6) require the Lipschitz continuity in terms of θ in $\Pi_\theta(x, dy)$ and $\pi_\theta(x)$ and in terms of x in $\Pi_\theta(x, dy)$. These conditions are usually satisfied in practice. (C.7) are moment and Lipschitz conditions on the functions $H(\theta, x)$ and $I(\theta, x)$. These should also be satisfied in practice. In particular, these conditions are satisfied in our example in Section 4.

We shall prove Theorem 1 in the Appendix.

6 Some Concluding Remarks.

We have proposed a stochastic approximation-based procedure for incomplete data estimation. The procedure incorporates naturally the Markov chain-based simulation procedure and can be used to resolve a wide class of incomplete data estimation problems. In addition, the proposed procedure is successfully applied to find the MLE in the structural measurement error logistic regression model, which, to the extend of our knowledge, has not been satisfyingly solved before.

Of course, the proposed stochastic approximation procedure is not limited to the estimation problem. For applications in adaptive control, our procedure serves as a correction to the of optimal choice of search direction suggested on page 115 of [9]. Their suggestion is equivalent, in our notation, to using $I(\theta, x) = -(\partial/\partial\theta)H(\theta, x)$, while in fact, a second term similar to that in (8) should be subtracted.

In addition, we have showed that under mild conditions, the proposed procedure converges. Our convergence theorem is based on Theorem 3.17 (or Corollary 3.18), Part II, of [9]. Even though our condition is easier to verify than those listed in [9], no effort is made to improve upon Theorem 3.17 of [9]. Our main objective is to show that the idea of stochastic approximation can successfully be applied to incomplete data estimation problems.

Appendix: Proofs of Theorem 1.

Theorem 1 is based on Theorem 3.17 (or Corollary 3.18), Part II, of [9] (page 304). First, we show that the updating step in (5) can be written in the form of equation (1.1.1) of [9] (page 213 or page 9). Using $\Gamma_k^{-1} - \Gamma_{k-1}^{-1} = -\Gamma_k^{-1}(\Gamma_k - \Gamma_{k-1})\Gamma_{k-1}^{-1}$ and the first expression of (5), we can write the second expression of (5) as

$$\theta_k = \theta_{k-1} + \gamma_k \Gamma_{k-1}^{-1} \bar{H}(\theta_{k-1}, \boldsymbol{X}_k) + \gamma_k^2 \rho_k(\theta_{k-1}, \Gamma_{k-1}, \boldsymbol{X}_k),$$

where

$$\rho_k(\theta, \Gamma, \boldsymbol{x}) = -\{I - \gamma_k \Gamma^{-1}(\bar{I}(\theta, \boldsymbol{x}) - \Gamma)\}\Gamma^{-1}(\bar{I}(\theta, \boldsymbol{x}) - \Gamma)\Gamma^{-1}\bar{H}(\theta, \boldsymbol{x}). \tag{15}$$

We see that expression (5) can be written in the form of expression (1.1.1) of [9] by viewing (θ, Γ) as a vector parameter.

We now show that under the assumptions of Theorem 1, the assumptions of Theorem 3.17 of [9] (namely (A.1), (A.2), (A.3), (A.4), (A$'$.5), (A$'$.6) and (A.7)) are satisfied.

(A.1) (p.213 of [9]). (A.1) is satisfied because of (C.1).

(A.2) (p.213 of [9]). We note that the random vector X_n in [9] is $(X_{k,1}, \ldots, X_{k,m})$ (n is k in our case). Thus, condition (A.2) is satisfied because of the way $(X_{k,1}, \ldots, X_{k,m})$ is simulated.

(A.3) (p.216 of [9]). We note that the parameter θ in [9] is (θ, Γ) in our case. The function $H(\theta, x)$ in [9] is $(\Gamma^{-1}\bar{H}(\theta, \boldsymbol{x}), \bar{I}(\theta, \boldsymbol{x}) - \Gamma)$, where the functions $\bar{H}(\theta, \boldsymbol{x})$ and $\bar{I}(\theta, \boldsymbol{x})$ are defined in (6). The function $\rho(\theta, \Gamma, \boldsymbol{x})$ is defined by (15). So condition (C.7) implies (A.3).

(A.4) (p.216 of [9]). To show that this condition holds, we need Lemma 2 below and Theorem 2.5 (page 259) of [9] The proof of Lemma 2 involves standard Markov chain arguments and is similar to the proof of a Harris convergent Markov chain must be geometrically convergent. See [14]. The proof of Lemma 2 is not provided.

Conditions (C.4) and (C.5) imply (j) of Lemma 2, which in turn implies condition (i) of Theorem 2.5 of [9] with $p_1 = q - 1$, $q_1 = q_2 = q$. Condition (C.3) implies condition (ii) of Theorem 2.5 of [9] with $m = q$. Again by Lemma 2, Conditions (C.4) and (C.5) imply (jj) of Lemma 2, which implies condition (iii) of Theorem 2.5 of [9] with $m = q$. Therefore, the conclusions of Theorem 2.5 of [9] hold with $p_1 = p_2 = q - 1$. Thus, (A.4) holds with $q_3 = q_4 = q$.

(A$'$.5) (p.290 of [9]). (i) of (A$'$.5) is guaranteed by (j) of Lemma 2. (i') of (A$'$.5) is guaranteed by (C.3). (ii) of (A$'$.5) is guaranteed by (C.6). (iii) of (A$'$.5) is implied by (C.5).

(A$'$.6) (p.301 of [9]). (C.2) implies (A$'$.6) with $\alpha = 2$.

(A.7) (p.233 of [9]). If we let D be the domain of attraction of $(\hat{\theta}, G(\hat{\theta}))$, then (A.7) holds on D. See the discussion on page 233 of [9].

Lemma 2 *Under the Assumptions (C.4) and (C.5), for any compact set $Q \in \Theta$*
(j) there exists a constant C_1' and $\rho_1 < 1$ such that for all $k \geq 1$, $\theta \in Q$ and $x \in E$

$$\int (1 + \|y\|^q) \left|\Pi_\theta^k(x, dy) - \pi_\theta(dy)\right| \leq C_1' \rho_1^k \left(1 + \|x\|^q\right);$$

(jj) there exists a constant C_2' and $\rho_2 < 1$, such that for all $k \geq 1$, $\theta, \theta' \in Q$ and $x \in E$

$$\int (1 + \|y\|^q) \left| \Pi_\theta^k(x, dy) - \pi_\theta(dy) - \Pi_{\theta'}^k(x, dy) + \pi_{\theta'}(dy) \right| \leq C_2' \rho_2^k \left| \theta - \theta' \right| (1 + \|x\|^q) \, .$$

REFERENCES

1. Lai, T. L. and Ying, Z. (1995) *Ann. Statist.* **22**, 1222-1255.

2. Schafer, D. W. (1987) *Biometrika* **74**, 385-391.

3. Gelfand, A. E. and Carlin, B. P. (1993) *Canad. J. Statist.* **21**, 303-311.

4. McCulloch, C. E. (1997) *J. Am. Statist. Assoc.*, **92**, 162-170.

5. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) *J. Chem. Phys.* **21**, 1087-1091.

6. Hastings, W. K. (1970) *Biometrika* **57**, 97-109.

7. Satten, G. A. (1996) *Biometrika* **83**, 355-370.

8. Robbins, H. and Monro, S. (1951) *Ann. Statist.* **22**, 400-407.

9. Benveniste, A., Métivier, M. and Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations* (Springer-Verlag, New York).

10. Louis, T. A. (1982) *J. R. Statist. Soc.* B, **44**, 98-130.

11. Tanner, M. A. (1996) *Tools for Statistical Inference, 3rd ed.* (Springer-Verlag, New York).

12. Clayton, D. (1991) in *Statistical Models for Longitudinal Studies of Health,* eds. Dwyer, J. H., Feinleib, M., Lippert, P. and Hoffmeister, H. (Oxford University Press, Oxford) 301-331.

13. Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models* (Chapman and Hall, New York).

14. Nummelin, E. (1984) *General Irreducible Markov Chain and Non-negative Operators* (Cambridge Press, Cambridge).