ESTIMATION FOLLOWING A ROBBINS-MONRO DESIGNED EXPERIMENT

by

Edward W. Frees

Department of Statistics and School of Business
1155 Observatory Drive
University of Wisconsin
Madison, Wisconsin 53706

and

David Ruppert

Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27514

January, 1986

Running Head: Robbins-Monro estimation

1980 AMS subject classification: Primary 62L20; Secondary 62L05; 62J05

Keywords and Phrases: stochastic approximation, stochastic regression, stochastic design

Abstract

Robbins and Monro (1951) initiated what is now a large literature on sequentially designed experiments for finding the root of an unknown regression function. Properties of the design have been extensively studied, especially the type and rate of convergence of the latest design point to the root. In this paper, we separate design and estimation considerations. A large class of designs are considered which converge to the root of a specified rate. For this class, a least-squares estimator of the root is introduced. Almost sure consistency and asymptotic normality of the estimator is proved. Further, the least-squares estimator is at least as efficient as the estimator typically suggested in these experiments.

§1. Introduction

Consider a general regression model,

$$Y_i = M(X_i) + \xi_i + \varepsilon_i, \qquad i=1,\ldots,n \tag{1.1}$$

where Y_i, X_i, ξ_i, ε_i are random vectors in R^p and M is a measurable function from R^p into R^p. At the design point X_i, the regression function M is measured with mean zero error ε_i and a small bias ξ_i (typically zero) which produces the observable output Y_i. The error ε_i is assumed to be a vector of martingale differences. The design can be chosen sequentially so that X_n depends on $X_1,\ldots,X_{n-1}$, $Y_1,\ldots,Y_{n-1}$. An important problem is to estimate the assumed unique root of $M(x) = \alpha$, say θ, where α is fixed and known. By subtracting α from the dependent variables Y_i, we can take α to be zero without loss of generality.

There are two distinct, though related, purposes in estimating θ. First, the data may be part of an experiment to study M, and the root of M is of particular importance. For example, in bioassay the LD_{50} is an important characteristic of the quantal response function. In this case, the actual sequence X_i is of interest only in so far as it affects the estimation of θ. Second, we may be attempting to control a process whose output is Y_i. If α is the optimal value of this output, then X_i should be as close to θ as possible for all i. Let l be a loss function such that $l(x) > 0$ when $x \neq 0$ and $l(0) = 0$. Then a suitable loss or cost for estimation is $l(\hat{\theta} - \theta)$ where $\hat{\theta}$ is an estimate of θ based on $X_1,\ldots,X_n$, $Y_1,\ldots,Y_n$. For control the cost would be

$$\sum_{i=1}^{n} l(X_i - \theta).$$

For both estimation and control Robbins and Monro (1951) proposed the stochastic approximation algorithm

$$X_{n+1} = X_n - a_n Y_n \tag{1.2}$$

for generating the design $\{X_n\}$ and using $\hat{\theta} = X_n$ to estimate θ. With this choice of $\hat{\theta}$, X_n should converge to θ as rapidly as possible, and there is no conflict between the goals of estimation and control. They assumed that $p = 1$ and $\xi_i \equiv 0$ and showed that X_n converges to θ in probability under suitable regularity conditions including

$$\sum_{i=1}^{\infty} a_i = \infty \quad \text{and} \quad \sum_{i=1}^{\infty} a_i^2 < \infty.$$

Convergence of the algorithm (1.2) was studied further by a host of researchers. Blum (1954a,b) introduced the multivariate version of (1.2) and showed that X_n converges to θ almost surely. Chung (1954) demonstrated the asymptotic normality of the univariate version of (1.2) and argued that $a_n = an^{-1}$, where "a" is a constant exceeding $1/(2\dot{M}(\theta))$, gives the best rate, $n^{-1/2}$, of convergence to normality. We will call such algorithms fixed "a" Robbins-Monro processes. Chung also showed that $a = 1/\dot{M}(\theta)$ minimizes the asymptotic variance of $n^{1/2}(X_n - \theta)$ and therefore, for purposes of estimation or control, if $\dot{M}(\theta)$ were known this choice of $\{X_n\}$ could be called asymptotically optimal. Of course, $\dot{M}(\theta)$ is rarely if ever known and asymptotic optimality is not possible in the case of fixed "a" processes.

Venter (1967) suggested using an estimate of $1/\dot{M}(\theta)$, say D_n, in place of a fixed "a". He showed that in this manner it is still possible to achieve the minimum asymptotic variance even without prior knowledge of $\dot{M}(\theta)$. Venter used a modification of the Robbins-Monro process, taking observations in pairs at $(X_{1,n}, X_{2,n}) = X_n \pm \gamma_n$ where γ_n converges to zero slowly. Call the corresponding Y observations $Y_{1,n}$ and $Y_{2,n}$. Using $\{Y_{1,n}, Y_{2,n}\}$ Venter could construct the estimate D_n of $\dot{M}(\theta)$. In algorithm (1.2) he replaced Y_n by the average of $Y_{1,n}$ and $Y_{2,n}$ and used $a_n = 1/(nD_n)$.

Lai and Robbins (1979, 1981) considered stochastic approximation specifically as a control procedure. They defined the "cost" of the algorithm to be $C_n = \sum_{i=1}^{n} (X_n - \theta)^2$. For fixed "a" procedures, $C_n/\log(n) \to \sigma^2$, the asymptotic variance of $n^{1/2}(X_n - \theta)$. If one could use $a_n = 1/(n\dot{M}(\theta))$, i.e., the optimal fixed "a" procedure, then lim $C_n/\log(n)$ would be minimized, and this procedure would be asymptotically optimal for both estimation and control. The Venter procedure is asymptotically optimal for estimation, but because the observations are taken not at X_n but at $X_n \pm \gamma_n$, the Venter procedure has a cost C_n that grows at an algebraic rate n^{λ}, $\lambda > 0$. Thus the cost C_n is of a higher order of magnitude for the Venter procedure than the fixed "a" procedures where $C_n = O(\log(n))$. Lai and Robbins (1979 and 1981) showed that estimation and control optimality need not conflict. They proposed using design (1.2) with $a_n = 1/(b_n n)$ where b_n is a least-squares estimate of $\dot{M}(\theta)$.

Specifically, Lai and Robbins suggested fitting the straight-line regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ to $\{Y_i, X_i\}$ and using $\hat{\beta}_1$ as an estimate of $\dot{M}(\theta)$. To ensure stability, a truncated version of $\hat{\beta}_1$ is used as b_n. This type of a_n sequence minimizes the asymptotic variance of $n^{1/2}(X_n - \theta)$ and the limit $C_n/\log(n)$. Lai and Robbins called such algorithms adaptive Robbins-Monro procedures.

In many situations, the cost C_n is irrelevant and only the accuracy of the final estimate of θ matters. Moreover, other aspects of M, perhaps $\dot{M}(\theta)$, may also be of interest. In this paper we show that in such cases the Lai and Robbins (1979, 1981) adaptive procedure has some undesirable features. Although it is asymptotically optimal as an estimate of θ, since X_n converges as rapidly as possible to θ, the design does not provide much information about $\dot{M}(\theta)$. For fixed "a" procedures with $a > 1/(2\dot{M}(\theta))$ we show that the

least-squares estimator of θ, $\hat{\theta} = -\hat{\beta}_0 / \hat{\beta}_1$, is asymptotically optimal as an estimate of θ even when X_n is not, and in addition that if X_n converges to θ at less than the optimal rate then $\hat{\beta}_1$ has a smaller asymptotic variance than in the Lai and Robbins design. This important result is apparently new even in the univariate case. It means that an optimal estimator of θ can be obtained from the easily implemented fixed "a" procedures even when $\dot{M}(\theta)$ is unknown. (Some knowledge of $\dot{M}(\theta)$ is necessary since $a > 1/(2\dot{M}(\theta))$ is required. However, usually an upper bound for $\dot{M}(\theta)$ is available.)

Our results are established also for the multivariate case, but the Lai and Robbins adaptive procedure has not been investigated there. It is obvious how the procedure itself can be extended, but to prove asymptotic properties of the extension would involve substantial technical difficulties. In the multivariate case our estimator $\hat{\theta}$ apparently is the only one known to be asymptotically optimal for estimation of θ and to have a cost C_n of order $\log(n)$.

Although the least-squares estimator is not as simple as X_n, it can be computed recursively, which may simplify computations. Indeed, the recursive nature of the Robbins-Monro algorithm is frequently mentioned in the engineering literature as a great advantage. This advantage is not lost if one uses the least-squares estimator, though one is then required to have two recursions, one for the Robbins-Monro algorithm and one for the least-squares estimator.

In this paper, the Robbins-Monro procedure is viewed as a method for constructing an experimental design. With this design, one can estimate the root θ of the regression function M and other characteristics of M, for example, the derivative at θ. We propose using a least-squares estimate of θ rather than the last observation of the Robbins-Monro process, as is traditional. By showing that the design and estimation can be separated profitably, we hope to

open a new area for investigation - how should a Robbins-Monro process be designed if only estimation, not control, is an issue and $\hat{\theta}$ can be any measurable function of $X_1,\ldots,X_n$, $Y_1,\ldots,Y_n$?

§2. Robbins-Monro Design Properties

We begin this section by giving some general notation and the assumptions about the regression model (1.1) that are used throughout the paper. We then cite some well-known facts about the properties of the design algorithm (1.2) and give a new result about the cost C_n which is a generalization of a result due to Lai and Robbins (1979, Theorem 2).

Assume that all random elements are defined on a fixed probability space (Ω, F, P) and that all relations between random elements are meant to hold almost surely, unless specified otherwise. For a random vector x_n in R^p, let x_{ni} be the i^{th} coordinate of x_n. Let A, A_1 be square, symmetric matrices. Denote the element in the i^{th} row and j^{th} column of A by $A^{(ij)}$. Define $\lambda_{min}(A)$ and $\lambda_{max}(A)$ to be, respectively, the smallest and largest eigenvalues of A. We write $A \geq A_1$ iff $A - A_1$ is nonnegative definite. Use $A = \text{diag}\,(a_i)$ to mean A is a diagonal matrix with a_i in the i^{th} row and column. Let vec(A) be the $p^2 \times 1$ vector built by stacking the column vectors of A. The symbol $\otimes$ denotes the usual Kronecker product of two matrices. The basic assumptions about the regression model (1.1) and design algorithm (1.2) are stated below.

A1. Let $\dot{M}(\theta)$ be a $p\times p$ matrix and $\eta > 0$. Suppose

$$M(x) = \dot{M}(\theta)(x-\theta) + O(|x-\theta|^{1+\eta}) \text{ as } x \to \theta.$$

A2. Assume that in (1.2), $a_n = n^{-1}D_n$ and assume $D_n \to D$, where D is such that $\lambda_{min}(D\dot{M}(\theta)) > 1/2$.

A3. Assume $n^{1/2}\,\xi_n \to 0$.

A4. Assume $X_n \to \theta$.

A5. There exists $\{F_n\}$, a nondecreasing sequence of sub σ-fields of F, so that:

(i) ε_n is a vector of martingale differences with respect to F_n,

(ii) for $\gamma > 2$, $\sup_n E(|\varepsilon_n|^\gamma | F_n) < \infty$ and,

(iii) for Σ, a positive definite p×p matrix, $E(\varepsilon_n \varepsilon_n' | F_n) \to \Sigma$.

A6. Assume D is symmetric and that $\dot{M}(\theta)$ is positive definite. Thus, for some orthogonal matrix P, $P(\dot{M}(\theta))^{1/2} D (\dot{M}(\theta))^{1/2} P' = \Lambda$, a diagonal matrix.

Remarks: Some implicit restrictions on the growth and smoothness of the regression function have been made by assuming A4, convergence of the algorithm. These restrictions are weak and we refer the reader to Ljung (1978) and Ruppert (1985) for recent discussions and further references. The most important assumption is the requirement in A6 that D and $\dot{M}(\theta)$ be symmetric. Here, D is a constant matrix and $\dot{M}(\theta)$ is the gradient matrix at the root θ. The symmetry requirement is not a restriction in the univariate case (p=1) and is important in the multivariate analysis since the basic approach is to rotate (1.1) to a vector of weakly correlated algorithms. Early multivariate stochastic approximation procedures assumed D to be some constant times the identity matrix, where the constant was chosen to (hopefully) satisfy A2. However, if D is taken to be $M(\theta)^{-1}$ or a constant multiple thereof, then the onerous assumption in A2 that $\lambda_{min}(D\dot{M}(\theta)) > 1/2$ is vacuous. Recent work in providing adaptive estimates D_n of D has been done by Lai and Robbins (1979, 1981) and Ruppert (1985). Our assumptions are general enough to cover both the classical (nonadaptive) and adaptive designs.

We now give some properties of the design (1.2).

Theorem 2.1

Consider the regression model (1.1) and stochastic design (1.2). Assume A1-A6. Then,

$$\sqrt{n}\,(X_n - \theta) \to_D N(0,\Sigma_1) \tag{2.2}$$

with $\Sigma_1 = (\dot{M}(\theta))^{-1/2}P'QP(\dot{M}(\theta))^{-1/2}$, where $Q^{(ij)} = (P(\dot{M}(\theta))^{1/2}D\ \Sigma\ D(\dot{M}(\theta))^{1/2}P')^{(ij)}/(\Lambda^{(ii)} + \Lambda^{(jj)} - 1)$. Further,

$$C_n/(\log n) \to \text{trace}\,(\Sigma_1). \tag{2.3}$$

Remarks: The relation (2.3) is a desirable property of the design and is proved in §4. The relation (2.2), originally due to Sacks (1958), together with the basic assumption A4, is used to justify X_n as an estimator of θ. Although X_n has desirable asymptotic properties, applied researchers (cf., Wetherill, 1966) have noted that it can perform poorly in finite samples. Heuristically, X_n is only the latest design point and it may be desirable to have an estimator which uses more averaging over other design points and responses. In the following section we introduce a least-squares estimator which uses more averaging. Further, unless $D = (\dot{M}(\theta))^{-1}$ this estimator is superior to X_n based on an asymptotic criterion.

§3. Least-Squares Estimation

The approach is to fit a hyperplane (line for p=1) through the observables (X_n,Y_n) and solve for the zero root of that hyperplane (line). Define X to be the $n \times (p+1)$ design matrix whose i^{th} row is given by $(1\ X_i')$. Let Y be the $n \times p$ matrix of responses whose i^{th} row is Y_i'. For sufficiently large n, define the $(p+1) \times p$ matrix of regression coefficients

$$\hat{\beta} = (X'X)^{-1} X'Y = (\hat{\beta}_0, \hat{\beta}_1)', \tag{3.1}$$

where $\hat{\beta}_0$ is a $p \times 1$ vector and $\hat{\beta}_1$ is a $p \times p$ matrix. The first result is a multivariate extension of Wei (1985, Theorem 2) and motivates the estimator of θ which we consider in the remainder of this section.

Theorem 3.1

Assume A1-A6. Then,

$$\hat{\beta}_0 \to - \dot{M}(\theta)\theta \tag{3.2}$$

and

$$\hat{\beta}_1 \to \dot{M}(\theta). \tag{3.3}$$

Thus, for sufficiently large n, we have that

$$\hat{\theta} = - \hat{\beta}_1^{-1} \hat{\beta}_0$$

is a consistent estimator of θ, i.e., $\hat{\theta} \to \theta$.

The estimator $\hat{\theta}$ is appealing because it is derived from the well-known regression coefficients matrix $\hat{\beta}$. It is well-defined for large n since $\hat{\beta}_1$ converges to $\dot{M}(\theta)$, a positive definite and hence invertible matrix. Although we justify the use of $\hat{\theta}$ by asymptotic efficiency, we are hopeful that $\hat{\theta}$ will have better finite sample properties than X_n even when D_n converges to $\dot{M}(\theta)^{-1}$ so that X_n is asymptotically efficient. As shown by Wu (1985), the Lai and Robbins adaptive procedure is not as successful in practice as its asymptotic properties would suggest. Davis (1971) has also shown that one must be quite careful about inferring finite sample properties of stochastic approximation algorithms from asymptotic results. Further, $\hat{\beta}$ can be expressed in a compact, recursive fashion. This is important in on-line estimation, loosely defined as a sampling situation where the data arrives quickly and it is important to have estimators readily available. For applications of on-line estimation and

stochastic approximation to electrical engineering, see Ljung and Söderström (1983).

Because $\hat{\theta}$ is based on the stochastic design matrix X, there is little hope of obtaining finite sample properties except under restrictive hypotheses on the distribution of ε. We do, however, have the following asymptotic result.

Theorem 3.2

Assume A1-A6. Then,

$$\sqrt{n}\,(\hat{\theta}-\theta) \to_D N(0,\Sigma_2)$$

where $\Sigma_2 = \dot{M}(\theta)^{-1}\,\Sigma\,\dot{M}(\theta)^{-1}$.

Remark: Following the usual terminology, we say that $\hat{\theta}$ is at least as efficient as X_n since $\Sigma_1 \geq \Sigma_2$. To see this, let b be an arbitrary vector in R^p and $b^* = P(\dot{M}(\theta))^{-1/2}\,b$. With $(\dot{M}(\theta))^{-1/2}\,P' = D(\dot{M}(\theta))^{1/2}\,P'\Lambda^{-1}$, we have

$$\begin{aligned} b'(\Sigma_1-\Sigma_2)b &= b^{*\prime}(Q - P(\dot{M}(\theta))^{-1/2}\,\Sigma(\dot{M}(\theta))^{-1/2}\,P')b^* \\ &= b^{*\prime}(Q - \Lambda^{-1}P(\dot{M}(\theta))^{1/2}\,D\Sigma D(\dot{M}(\theta))^{1/2}\,P'\Lambda^{-1})b^* \\ &= \sum_{i,j} b_i^* b_j^* (P(\dot{M}(\theta))^{1/2}\,D\Sigma D(\dot{M}(\theta))^{1/2}\,P')^{(ij)} \\ &\qquad \left((\Lambda^{(ii)} + \Lambda^{(jj)} - 1)^{-1} - (\Lambda^{(ii)}\,\Lambda^{(jj)})^{-1}\right) \\ &= \tilde{b}'\,Q\,\tilde{b} \geq 0 \end{aligned}$$

with $\tilde{b}_i = (1 - 1/\Lambda^{(ii)})b_i^*$, since Q is positive definite.

Further remarks: When $\dot{M}(\theta)$ is symmetric, Σ_2 is the asymptotic variance-covariance matrix of the Ruppert's (1985) modified Robbins-Monro procedure. In the univariate case we have $\Sigma_2 = \Sigma/(\dot{M}(\theta))^2$. Thus, if the design (1.2) is used with $D_n = D = a$, Σ_2 is invariant to the choice of the design parameter a

(subject to A2) and equals the asymptotic variance of the efficient procedure of Lai and Robbins (1979 and 1981). We note, however, that the cost of the design is not invariant to the choice of a. By Theorem 2.1, the cost is of order (log n) $a^2\Sigma/(2a\dot{M}(\theta)-1)$. Asymptotically, this is at least as large as Lai and Robbin's cost which is of order (log n) $\Sigma/(\dot{M}(\theta))^2$. We have equality if $a = (\dot{M}(\theta))^{-1}$.

Theorem 3.2 provides information about the rate of convergence of $\hat{\theta}$ to θ and gives a rationale for preferring $\hat{\theta}$ to X_n. It is also helpful in providing confidence regions for θ in conjunction with the following

Theorem 3.3.

Assume A1-A6 and define

$$\hat{\Sigma} = n^{-1}(Y-X\hat{\beta})'(Y-X\hat{\beta}). \tag{3.5}$$

Then $\hat{\Sigma} \to \Sigma$.

Corollary 3.1

Assume A1-A6. Then, for $c \geq 0$, $P(n(\hat{\theta}-\theta)'\hat{\beta}_1\hat{\Sigma}^{-1}\hat{\beta}_1(\hat{\theta}-\theta) \leq c) \to P(\chi_p^2 \leq c)$ where χ_p^2 is a chi-square random variable with p degrees of freedom.

For $\alpha^* \in (0,1)$, we can choose c so that $\alpha^* = P(\chi_p^2 \leq c)$. Thus, Corollary 3.1 provides an approximate confidence ellipsoid for θ.

From Theorem 3.1, we have that $\hat{\beta}_1$ is a consistent estimator of the gradient matrix at the root, $\dot{M}(\theta)$. We have further information about that rate of convergence contained in the following

Theorem 3.4.

Assume A1-A6. Then

$$(\log n)^{1/2}\ \mathrm{vec}(\hat{\beta}_1-\dot{M}(\theta)) \to_D N(0,(\Sigma \otimes (P'QP)^{-1/2}\dot{M}(\theta)(P'QP)^{-1/2}))$$

where Q is defined in Theorem 2.1.

Remarks: In the univariate case, the asymptotic variance of the standardized slope estimate is $\Sigma\dot{M}(\theta)Q^{-1} = (2D\dot{M}(\theta)-1)/D^2$. Typically, the choice $D = (\dot{M}(\theta))^{-1}$ is said to be the optimal choice since it minimizes the asymptotic variance of the estimator X_n and standardized asymptotic cost $\Sigma_1 = D^2\Sigma/(2D\dot{M}(\theta)-1)$. It is interesting to note that if the researcher is also interested in estimating the slope at the root that this is the worst possible choice of D in terms of the estimator's asymptotic variance. Heuristically, we think of $\lim_{n\to\infty} \sum_{k=1}^{n}(X_k-\theta)^2/(\log n) = \Sigma_1$ as a measure of how spread out the design points X_n are. The closer the parameter D is to $(\dot{M}(\theta))^{-1}$, the smaller the spread of the design. In this light, it is reasonable that as D moves farther away from $(\dot{M}(\theta))^{-1}$ the design becomes more spread out and thus our estimate of the slope at the root improves in terms of its asymptotic variance. We comment further on this aspect of the design in §6. We finally remark that with $D = (\dot{M}(\theta))^{-1}$ the asymptotic variance of $\hat{\beta}_1$ is the same as Lai and Robbin's (1981, Theorem 4) estimator.

§4. Design Properties

In this section we establish some properties of the design which are then used to prove Theorem 2.1 and are also needed in §5. This section is divided into five lemmas, the first two being technical devices. The third lemma shows how to approximate $X_n-\theta$ sufficiently closely by a weighted sum of martingale difference errors, a multivariate generalization of Theorem 3 of Lai and Robbins (1979). The fourth lemma establishes an order of magnitude for $X_n-\theta$. The fifth and final lemma establishes the proof of Theorem 2.1 as an immediate corollary.

We now briefly collect some additional facts which we use repeatedly in §4 and §5. For column vectors we use the usual Euclidean norm. Let A be a

$p \times p$ matrix. We use the norm $|A| = (\sum_{i=1}^{p} \sum_{j=1}^{p} (A^{(ij)})^2)^{1/2}$. Define $\exp(A) = \sum_{i=0}^{\infty} A^i/i!$ and for $t > 0$, define $t^A = \exp(\log(t)A)$. A sequence $\{f_n\}$ of real numbers is said to be slowly varying if $f_{[tn]}/f_n \to 1$ for each $t > 0$. The following are properties of slowly varying sequences, contained in Theorems 3, 4 and 6 of Bojanic and Seneta (1973). If $\{f_n\}$ is slowly varying, then $|f_n|$ is bounded above and below by $C_1 n^{\varepsilon}$ and $C_2 n^{-\varepsilon}$ for every $\varepsilon > 0$, where C_1 and C_2 are positive constants. If $f_n^{-1}(f_{n+1} - f_n) = o(n^{-1})$, then $\{f_n\}$ is slowly varying. For $g > -1$ and $\{f_n\}$ slowly varying, we have $\sum_{k=1}^{n} k^g f_k \sim n^{g+1} f_n/(g+1)$.

The following lemma is a useful multivariate generalization of the third property.

Lemma 4.1

Let $\{A_n\}$ be a sequence of positive $p \times p$ matrices such that $A_n^{-1}(A_{n+1} - A_n) = o(n^{-1})$. Then $\{|A_n|\}$ is a sequence of slowly varying real numbers. Further, let B be a $p \times p$ matrix so that $\lambda_{min}(B) > 0$ and $\{C_{pn}\}$ a sequence of $p \times p$ matrices such that $C_{pn} \to C_p$. Then

$$\sum_{k=1}^{n} A_k \, k^{B-I} \, C_{pk} \sim A_n \, n^B \, B^{-1} \, C_p. \tag{4.1}$$

Proof:

To show that $\{|A_n|\}$ is slowly varying, we have

$$n \, |A_n^{-1}| \, (|A_n| - |A_{n+1}|) \le n \, (|I| - |A_n^{-1} A_{n+1}|)$$

$$\le n \, |I - A_n^{-1} A_{n+1}| \to 0.$$

The proof of the reverse inequality is similar.

To show (4.1), via a Taylor-series expansion, we have $n^B = (n-1)^B + B \, n^{B-I} + O(n^{B-2I})$ and thus,

$$n^{B-I} = (n^B - (n-1)^B) \, B^{-1} + O(n^{B-2I}).$$

Now, reversing the order of summation,

$$\sum_{k=1}^{n} A_k\, k^{B-I}\, C_{pk} = \sum_{k=1}^{n} A_k\, k^{B-I}\, C_p + \sum_{k=1}^{n} o(A_k\, k^{B-I})$$

$$= \sum_{k=1}^{n} A_k\, (k^B - (k-1)^B)\, B^{-1} C_p + \sum_{k=1}^{n} o(A_k\, k^{B-I})$$

$$= A_n\, n^B\, B^{-1}\, C_p - \sum_{k=1}^{n} (A_k - A_{k-1})(k-1)^B\, B^{-1} C_p$$

$$+ \sum_{k=1}^{n} o(A_k\, k^{B-I})$$

$$= A_n\, n^B\, B^{-1}\, C_p + \sum_{k=1}^{n} o(A_k\, k^{B-I}).$$

We get the result via an application of the Toeplitz lemma. ‡

Lemma 4.2

Let $\{U_n\}$ be a sequence of p×p matrices such that U_n is F_{n-1}-measurable and $\{\varepsilon_n\}$ satisfy A5 (i,ii). Define $s_n^2 = \sum_{i=1}^{n} |U_i|^2$. If $|U_n|^2 = o(s_n^{2c})$ for some $0 < c < 1$, then,

$$\sum_{i=1}^{n} U_i \varepsilon_i = O(s_n (\log_2 s_n^2)^{1/2}).$$

The proof of Lemma 4.2 is a straightforward multivariate extension of a special case of Lemma 2 of Wei (1985).

Lemma 4.3

Assume A1-A6. Define $C = (\dot{M}(\theta))^{1/2}\, D (\dot{M}(\theta))^{1/2}$ and $\delta_k = (\dot{M}(\theta))^{1/2}\, D_k$. Then

$$X_{n+1} - \theta = -(\dot{M}(\theta))^{1/2}\, n^{-C}\, \tau_n^{-1} \left(\sum_{k=1}^{n} \tau_k\, k^{C-I}\, \delta_k\, \varepsilon_k + \rho \right) + o(n^{-1/2}),$$

where ρ is a random vector and $\{\tau_n\}$ is a sequence of random matrices such that

$$\tau_n^{-1}(\tau_{n+1} - \tau_n) = o(n^{-1}).$$

Proof:

From A1, let $d_n = \dot{M}(\theta) + O(|X_n-\theta|^{\eta})$. Define $\Phi_n = (\dot{M}(\theta))^{1/2}(X_n-\theta)$ and premultiply (1.2) by $(\dot{M}(\theta))^{1/2}$ to get

$$\Phi_{n+1} = (I - n^{-1}(\dot{M}(\theta))^{1/2} D_n d_n (\dot{M}(\theta))^{-1/2}) \Phi_n - n^{-1} \delta_n(\xi_n + \varepsilon_n).$$

Iterating the algorithm, we have

$$\Phi_{n+1} = \beta_{m-1,n} \Phi_m - \sum_{k=m}^{n} \beta_{k,n} k^{-1} \delta_k(\xi_k + \varepsilon_k) \tag{4.2}$$

where $\beta_{k,n} = (I - n^{-1} B_n) \ldots (I - (k+1)^{-1} B_{k+1})$,

$$B_k = \begin{cases} (\dot{M}(\theta))^{1/2} D_k d_k (\dot{M}(\theta))^{-1/2} & \text{if } \lambda_{min}[(\dot{M}(\theta))^{1/2} D_k d_k (\dot{M}(\theta))^{-1/2}] > 0 \\ 1/2\ I & \text{otherwise} \end{cases}$$

and m is a positive random integer chosen so that $B_n = (\dot{M}(\theta))^{1/2} D_n d_n (\dot{M}(\theta))^{-1/2}$ for $n \geq m$. By A2 and A4, $B_n \to C$, $\lambda_{min}[(\dot{M}(\theta))^{1/2} D_n d_n (\dot{M}(\theta))^{-1/2}] > 1/2$ for all sufficiently large n and thus m is finite almost surely.

Define $\tau_n = \beta_{0,n}^{-1} n^{-C}$ and note that

$$\tau_n^{-1}(\tau_{n+1} - \tau_n) = o(n^{-1}). \tag{4.3}$$

To see (4.3), by A1 and A2, for large n,

$$\begin{aligned} \tau_n^{-1}(\tau_{n+1} - \tau_n) &= n^C(I-(n+1)^{-1}B_{n+1})^{-1}(n+1)^{-C} - I \\ &= n^C(I + n^{-1}C(1+o(1))(n+1)^{-C} - I \\ &= (n/n+1)^C + n^{-1}n^C C(n+1)^{-C}(1+o(1)) - I = o(n^{-1}). \end{aligned}$$

With $\beta_{k,n} = n^{-C}\,\tau_n^{-1}\tau_k\,k^C$, from (4.2) we have

$$\Phi_{n+1} = n^{-C}\,\tau_n^{-1}\tau_{m-1}(m-1)^C\,\Phi_m - n^{-C}\,\tau_n^{-1}\sum_{k=m}^{n}\tau_k\,k^{C-I}\delta_k(\xi_k + \varepsilon_k). \tag{4.4}$$

By Lemma 4.1 and (4.3), $\tau_n^{-1} = O(n^{\varepsilon})$. Thus, from A2,

$$n^{-C}\,\tau_n^{-1} = o(n^{-1/2}).$$

By A1-A3, (4.3) and Lemma 4.1, it can be shown that

$$n^{-C}\,\tau_n^{-1}\sum_{k=m}^{n}\tau_k\,k^{C-I}\,\delta_k\,\xi_k = n^{-C}\,\tau_n^{-1}\sum_{k=m}^{n}\tau_k\,k^{C-I}\,o(k^{1/2}) = o(n^{-1/2}).$$

Thus, from (4.4), we have

$$\Phi_{n+1} = -n^{-C}\,\tau_n^{-1}\sum_{k=m}^{n}\tau_k\,k^{C-I}\,\delta_k\varepsilon_k + o(n^{-1/2})$$

which is sufficient for the result. ♯

Lemma 4.4

Assume A1-A6. Then

$$X_{n+1} - \theta = O((\log_2 n/n)^{1/2}) \tag{4.5}$$

and

$$\sum_{k=1}^{n} k^{-1/2}(X_{k+1} - \theta) = o(\log n). \tag{4.6}$$

Proof:

We intend to show that

$$(\dot{M}(\theta))^{1/2}(X_{n+1} - \theta) = n^{-C}\,S_1(n) + o((\log_2 n/n)^{1/2}). \tag{4.7}$$

where $S_1(n) = \sum_{k=1}^{n} k^{C-I}\,\delta_k\varepsilon_k$. To see that this is sufficient for (4.5), by Lemmas 4.1 and 4.2, we have

$$n^{-C}S_1(n) = O(n^{-C}(\sum_{k=1}^{n} | k^{C-I}\ \delta_k|^2)^{1/2}\ (\log_2 \sum_{k=1}^{n} |k^{C-I}\ \delta_k|^2)^{1/2})$$

$$= O((\log_2 n/n)^{1/2}). \tag{4.8}$$

To prove (4.7), by reversing the order of summation,

$$S_n = \sum_{k=1}^{n} \tau_k\ k^{C-I}\ \delta_k\varepsilon_k = \sum_{k} \tau_k(S_1(k) - S_1(k-1))$$

$$= \tau_n S_1(n) - \sum_{k=1}^{n} (\tau_k - \tau_{k-1})S_1(k-1).$$

$$= \tau_n S_1(n) + O(1) + o(\sum_{k} k^{-1}\ \tau_{k-1}S_1(k-1))$$

$$= \tau_n S_1(n) + O(1) + o(\sum_{k} \tau_{k-1}k^{C-3/2I}(\log_2 k)^{1/2})$$

$$= \tau_n S_1(n) + O(1) + o(\tau_n\ n^{C-1/2I}(\log_2 n)^{1/2}) \tag{4.9}$$

from (4.3), (4.8) and Lemma 4.1. This and Lemma 4.3 are sufficient for (4.7) and hence (4.5).

To prove (4.6), first define $a_1(n) = \sum_{k=n}^{\infty} k^{-C - 1/2I}$. Via Taylor-series expansion, it can be shown that $(C - 1/2I)\ n^{C - 1/2I}\ a_1(n) \to I$. Now, by reversing the order of summation,

$$\sum_{k=1}^{n} k^{-1/2}\ k^{-C}\ \tau_k^{-1}\ S_k$$

$$= \sum_{k=1}^{n} (a_1(k) - a_1(k+1))\ \tau_k^{-1}\ S_k$$

$$= a_1(1)\ \tau_1^{-1}\ S_1 - a_1(n+1)\ \tau_n^{-1}\ S_n$$

$$+ \sum_{k=1}^{n} a_1(k+1)(\tau_{k+1}^{-1} - \tau_k^{-1})\ S_{k+1} + \sum_{k=1}^{n} a_1(k+1)\ \tau_k^{-1}\ (S_{k+1} - S_k)$$

$$= O((\log_2 n)^{1/2}) + \sum_{k=1}^{n} o(k^{-C-1/2I}\, \tau_k^{-1}\, S_{k+1})$$

$$+ \sum_{k=1}^{n} a_1(k+1)\, \tau_k^{-1}\, \tau_{k+1}\, (k+1)^{C-I}\, \delta_{k+1}\varepsilon_{k+1}$$

$$= O((\log_2 n)^{1/2}) + o(\sum_{k=1}^{n} k^{-C-1/2I}\, \tau_k^{-1}\, S_k)$$

$$+ (\sum_{k=1}^{n} (k+1)^{-1/2}\, \delta_{k+1}\, \varepsilon_{k+1})(1 + o(1))$$

$$= O((\log_2 n)^{1/2} + (\log n \log_3 n)^{1/2}) + o(\sum_{k=1}^{n} k^{-C-1/2I}\, \tau_k^{-1}\, S_k).$$

This and Lemma 4.3 are sufficient for the proof. #

Lemma 4.5

Assume A1-A6 and define $T_n = (\dot{M}(\theta))^{1/2} (\sum_{k=1}^{n} (X_k-\theta)(X_k-\theta)')(\dot{M}(\theta))^{1/2}$. Then, $T_n/(\log n) \to P'QP$ where Q is defined in Theorem 2.1.

Proof:

By Lemma 4.3 and (4.6),

$$T_n = T_1(n) + o(\log n), \tag{4.10}$$

where

$$T_1(n) = \sum_{k=1}^{n} k^{-C}\, \tau_k^{-1}\, S_k S_k'\, (k^{-C}\, \tau_k^{-1})' \tag{4.11}$$

and

$$S_n = \sum_{k=1}^{n} \tau_k\, k^{C-I}\, \delta_k \varepsilon_k.$$

Now, use $S_0 = 0$ and let $S_n S_n' = S_2(n) + S_3(n) + S_3(n)'$, where

$$S_2(n) = \sum_{k=1}^{n} \tau_k\, k^{C-I}\, \delta_k \varepsilon_k \varepsilon_k'\, (\tau_k\, k^{C-I}\, \delta_k)'$$

and

$$S_3(n) = \sum_{k=1}^{n} \tau_k\, k^{C-I}\, \delta_k \varepsilon_k S_{k-1}'.$$

Define $a_2(n) = \sum_{k=n}^{\infty} k^{-2C}$. As in the proof of Lemma 4.4, we have

$$\lim_{n\to\infty} (2C - 1)\, n^{2C-I}\, a_2(n) = I. \tag{4.12}$$

Thus, since $|\sum_{k=1}^{n} k^{-C}\, \tau_k^{-1}\, S_3(k)(k^{-C}\tau_k^{-1})'| = |\sum_{k=1}^{n} k^{-2C}\tau_k^{-1}\, S_3(k)(\tau_k^{-1})'$,

we consider

$$\begin{aligned}
&\sum_{k=1}^{n} k^{-2C}\tau_k^{-1}\, S_3(k)(\tau_k^{-1})' \\
&= \sum_{k=1}^{n} (a_2(k) - a_2(k+1))\, \tau_k^{-1}\, S_3(k)(\tau_k^{-1})' \\
&= a_2(1)\, \tau_1^{-1}\, S_3(1)(\tau_1^{-1})' - a_2(n+1)\, \tau_{n+1}^{-1}\, S_3(n+1)(\tau_{n+1}^{-1})' \\
&\quad + \sum_{k=1}^{n} a_2(k+1)(\tau_{k+1}^{-1}\, S_3(k+1)(\tau_{k+1}^{-1})' - \tau_k^{-1} S_3(k)(\tau_k^{-1})') \\
&= O(1) + O(n^{-2C+I}\, \tau_{n+1}^{-1}\, S_3(n+1)(\tau_{n+1}^{-1})') \\
&\quad + \sum_{k=1}^{n} a_2(k+1)\{(\tau_{k+1}^{-1} - \tau_k^{-1})\, S_3(k+1)(\tau_k^{-1})' \\
&\quad + \tau_k^{-1}(S_3(k+1) - S_3(k))(\tau_k^{-1})' \quad + \tau_{k+1}^{-1}\, S_3(k+1)(\tau_{k+1}^{-1} - \tau_k^{-1})'\} \\
&= O(1) + O(n^{-2C+I}\, \tau_{n+1}^{-1}\, S_3(n+1)(\tau_{n+1}^{-1})') \\
&\quad + o(\sum_{k=1}^{n} k^{-2C}\, \tau_k^{-1}\, S_3(k)(\tau_k^{-1})') \\
&\quad + \sum_{k=1}^{n} a_2(k+1)\, \tau_k^{-1}\, \tau_{k+1}(k+1)^{C-I}\, \delta_{k+1}\, \varepsilon_{k+1}\, S_k'\, (\tau_k^{-1})'.
\end{aligned} \tag{4.13}$$

Now $\varepsilon_k\, S_{k-1}'\, \tau_{k-1}'$ is a $p\times p$ matrix of martingale differences times a row vector transform. Thus, a column by column application of Lemma 4.2 gives

$$\begin{aligned}
&\sum_{k=2}^{n+1} a_2(k)\, \tau_{k-1}^{-1}\, \tau_k\, k^{C-I}\, \delta_k\, \varepsilon_k\, S_{k-1}'\, (\tau_{k-1}^{-1})' \\
&= O(t_n\, (\log_2 t_n^2)^{1/2}) = o(t_n^2) + O(1)
\end{aligned} \tag{4.14}$$

where $t_n^2 = \sum_{k=1}^{n} |a_2(k)\ \tau_k^{-1}\ k^{C-I}\ \delta_k|^2 |S_{k-1}|^2$. From (4.7), (4.9), (4.12) and Lemma 4.1, we have

$$t_n^2 = O(\sum_{k=1}^{n} |\tau_k^{-1}\ k^{-C}|^2 |S_{k-1}|^2) = O(T_1(n)). \tag{4.15}$$

Similarly,

$$S_3(n) = O((\sum_{k=1}^{n} |\tau_k\ k^{C-I}\ \delta_k|^2 |S_{k-1}|^2)^{1/2} (\log_2 \sum_{k=1}^{n} |\tau_k\ k^{C-I}\ \delta_k|^2 |S_{k-1}|^2)^{1/2})$$

$$= O(|\tau_n|^2\ n^{2C-I} (\log_2 n)).$$

Thus, using this, (4.12), (4.14), and (4.15) in (4.13), we get

$$\sum_{k=1}^{n} k^{-C}\ \tau_k^{-1}\ S_3(k)(k^{-C}\ \tau_k^{-1})' = o(T_1(n)) + O(\log_2 n).$$

Thus, using (4.10) and (4.11), we only need to show

$$(\log n)^{-1} \sum_{k=1}^{n} k^{-C}\ \tau_k^{-1}\ S_2(n)\ (k^{-C}\ \tau_k^{-1})' \to P'QP. \tag{4.16}$$

Define

$$S_4(n) = \sum_{k=1}^{n} \tau_k\ k^{C-I}\ \delta_k\ E(\varepsilon_k \varepsilon_k' | F_{k-1})(\tau_k\ k^{C-I}\ \delta_k)'$$

and

$$S_5(n) = \sum_{k=1}^{n} \tau_k\ k^{C-I}\ \delta_k\ \Sigma\ (\tau_k\ k^{C-I}\ \delta_k)'.$$

Using Lemmas 4.1 and 4.2, it can be shown that

$$(\log n)^{-1} \sum_{k=1}^{n} k^{-C}\ \tau_k^{-1}\ (S_2(k) - S_4(k))\ (k^{-C}\ \tau_k^{-1})' \to 0$$

and

$$(\log n)^{-1} \sum_{k=1}^{n} k^{-C}\ \tau_k^{-1}\ (S_4(k) - S_5(k))(k^{-C}\ \tau_k^{-1})' \to 0.$$

Thus, sufficient for (4.16) is to show

$$(\log n)^{-1} \sum_{k=1}^{n} P\ k^{-C}\ \tau_k^{-1}\ S_5(k)(P\ k^{-C}\ \tau_k^{-1})' \to Q. \tag{4.17}$$

Since $\delta_k \to (\dot{M}(\theta))^{1/2} D$, it can be shown that

$$S_5(n) \sim \sum_{k=1}^{n} \tau_k k^{C-I} (\dot{M}(\theta))^{1/2} D \Sigma D (\dot{M}(\theta))^{1/2} (\tau_k k^{C-I})'$$

$$= \sum_{k=1}^{n} \tau_k P' k^{\Lambda-I} Q_1 k^{\Lambda-I} (\tau_k^{-1} P')'.$$

Here, we define $Q_1 = P(\dot{M}(\theta))^{1/2} D \Sigma D(\dot{M}(\theta))^{1/2} P'$ and note that from A6 $P k^C P' = k^\Lambda$. Define $\lambda_i = \Lambda^{(ii)}$ and note that the $(i,j)^{th}$ element of $k^{\Lambda-I} Q_1 k^{\Lambda-I}$ is $k^{\lambda_i+\lambda_j-2} Q_1^{(ij)}$. Similarly to Lemma 4.1, via a partial summation technique, it can be shown that

$$S_5(n) \sim (\tau_n P') \left(\frac{n^{\lambda_i+\lambda_j-1} Q_1^{(ij)}}{\lambda_i + \lambda_j - 1}\right)_{(ij)} (\tau_n P')'$$

$$= n^{-1} \tau_n P' n^\Lambda Q n^\Lambda (\tau_n P')'.$$

This is sufficient for (4.17) and hence the result. ♯

§5. Proof of Section 3 Results

Before giving the proofs of the §3 results, we first provide some additional notation. Define the rotation matrix

$$R = \begin{pmatrix} 1 & -\theta'(\dot{M}(\theta))^{1/2} \\ 0 & (\dot{M}(\theta))^{1/2} \end{pmatrix}$$

and $Z = XR$, the rotated design matrix. Let $\underset{\sim}{M}(X)$, ξ and ε be $n \times p$ matrices whose i^{th} row is given by, respectively, $M(X_i)'$, ξ_i' and ε_i'. Define $\beta = (-\dot{M}(\theta)\theta, \dot{M}(\theta))'$. It is easy to check that

$$R^{-1} = \begin{pmatrix} 1 & \theta' \\ 0 & (\dot{M}(\theta))^{-1/2} \end{pmatrix}$$

and, with (3.1), we have

$$R^{-1}\hat{\beta} = (\hat{\beta}_0 + \hat{\beta}_1\theta,\ (\dot{M}(\theta))^{-1/2}\ \hat{\beta}_1)' = (Z'Z)^{-1}Z'Y. \tag{5.1}$$

We will also need

$$H = \begin{pmatrix} n & 0 \\ 0 & (\log n)(P'QP) \end{pmatrix}.$$

We preface the proof of Theorem 3.1 with a preparatory lemma.

Lemma 5.1

Assume A1-A6. Then

$$Z'\underset{\sim}{M}(X) = Z'ZR^{-1}\beta + o(H^{1/2}).$$

Proof: First note that $R^{-1}\beta = (0,\ (\dot{M}(\theta))^{1/2})'$. Now, for the first row of $Z'\underset{\sim}{M}(X)$, we have by A1 and (4.5),

$$\begin{aligned}\sum_{k=1}^{n} M(X_k)' &= \sum_{k=1}^{n} (X_k - \theta)'\dot{M}(\theta) + O(\sum_{k=1}^{n} |X_k - \theta|^{1+\eta}) \\ &= \sum_{k=1}^{n} (X_k - \theta)'\dot{M}(\theta) + O(\sum_{k=1}^{n} (\log_2 k/k)^{(1+\eta)/2}) \\ &= \sum_{k=1}^{n} (X_k - \theta)'\dot{M}(\theta) + o(n^{-1/2}),\end{aligned}$$

since $\eta > 0$.

For the remaining p rows of Z'M(X), by A1 and (4.5),

$$\begin{aligned}(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n} (X_k-\theta)M(X_k)' &= (\dot{M}(\theta))^{1/2} \sum_{k=1}^{n} (X_k-\theta)(X_k-\theta)'\dot{M}(\theta) + O(\sum_{k=1}^{n} |X_k-\theta|^{2(1+\eta)}) \\ &= (\dot{M}(\theta))^{1/2}(\sum_{k=1}^{n} (X_k-\theta)(X_k-\theta)')(\dot{M}(\theta))^{1/2}(\dot{M}(\theta))^{1/2} + O(1)\end{aligned}$$

which is sufficient for the lemma. #

Proof of Theorem 3.1: By Lemmas 4.4 and 4.5, we have

$$H^{-1/2}Z'ZH^{-1/2} \to I. \tag{5.2}$$

Now, with (5.1) and Lemma 5.1,

$$R^{-1}\hat{\beta} = (Z'Z)^{-1}Z'(\underset{\sim}{M}(X) + \xi + \varepsilon)$$

$$= R^{-1}\beta + o((Z'Z)^{-1}H^{1/2}) + (Z'Z)^{-1}Z'(\xi + \varepsilon).$$

Thus, with (5.2),

$$R^{-1}(\hat{\beta} - \beta) = (Z'Z)^{-1}Z'(\xi + \varepsilon) + o(H^{-1/2}). \tag{5.3}$$

The proof of the theorem will be complete when we show

$$H^{1/2}(Z'Z)^{-1}Z'\xi \to 0 \tag{5.4}$$

and

$$(Z'Z)^{-1}Z'\varepsilon \to 0. \tag{5.5}$$

Equation (5.4) is stronger than we need now but it will be useful in its present form later. To prove (5.4), by (5.2) we need only show $H^{-1/2}Z'\xi \to 0$. By A3, the first row of $Z'\xi$ is $\sum_{k=1}^{n} \xi_k' = o(n^{1/2})$. The remaining p rows of $Z'\xi$ are

$$(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n} (X_k - \theta)\, \xi_k = o(\log n)$$

by A3 and Lemma 4.4. This is sufficient for (5.4).

To prove (5.5), first note that

$$(Z'Z)^{-1}Z'\varepsilon = H^{-1/2}(I + o(1))H^{-1/2}Z'\varepsilon \tag{5.6}$$

by (5.2). Now, the first row of $Z'\varepsilon$ is $\sum_{k=1}^{n}\varepsilon_k' = o(n)$ by the strong law of large numbers for martingales. The remaining p rows of $Z'\varepsilon$ are

$$(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n} (X_k - \theta)\, \varepsilon_k' = O\Big(\Big(\sum_{k=1}^{n} |X_k - \theta|^2 \log_2\Big(\sum_{k=1}^{n} |X_k - \theta|^2\Big)\Big)^{1/2}\Big)$$

$$= O((\log n \log_3 n)^{1/2})$$

by a column-wise application of Lemma 4.2 and Lemma 4.3. This and (5.6) are sufficient for (5.5) and hence the theorem. ♯

Proof of Theorem 3.2: By (5.3), (5.4), and (5.6), we have

$$H^{1/2}R^{-1}(\hat{\beta} - \beta) = (I + o(1))H^{-1/2}Z'\varepsilon + o(1). \tag{5.7}$$

The first row of (5.7) is

$$\sqrt{n}\,(\hat{\beta}_0' + \theta'\hat{\beta}_1) = (1 + o(1))n^{-1/2} \sum_{k=1}^{n} \varepsilon_k' + o(1).$$

Taking transposes and the usual central limit theorems for vectors of martingale differences give

$$\sqrt{n}\,(\hat{\beta}_0 + \hat{\beta}_1\theta) \to_D N(0,\Sigma)$$

which, together with Theorem 3.1, is sufficient for the theorem. ♯

Proof of Theorem 3.3: Define $\Delta = \underset{\sim}{M}(X) - X\beta + \xi$ and $C = I - X(X'X)^{-1}X' = I - Z(Z'Z)^{-1}Z'$, where $CX = CZ = 0$ and C is idempotent. Now,

$$Y - X\hat{\beta} = CY = C(X\beta + \Delta + \varepsilon) = C(\Delta + \varepsilon).$$

By A5(iii) and Theorem 5 of Chow (1965), we have $n^{-1}\varepsilon'\varepsilon \to \Sigma$. Thus, we wish to show

$$\hat{\Sigma} - n^{-1}\varepsilon'\varepsilon = n^{-1}((\Delta + \varepsilon)'C(\Delta + \varepsilon) - \varepsilon'\varepsilon) \to 0. \tag{5.8}$$

By Lemma 5.1, (5.2), (5.4) and (5.5), $(Z'Z)^{-1}Z'(\Delta + \varepsilon) \to 0$. Thus, $n^{-1}(\Delta + \varepsilon)'Z(Z'Z)^{-1}Z'(\Delta + \varepsilon) \to 0$. To prove (5.8), we only need to show

$$n^{-1}(\Delta'\Delta + \Delta'\varepsilon + \varepsilon'\Delta) \to 0$$

which follows by similar arguments. ♯

Proof of Theorem 3.4: From the bottom of p rows of (5.7) and (5.1), we have

$$(\log n)^{1/2}(P'QP)^{1/2}(\dot{M}(\theta))^{-1/2}(\hat{\beta}_1 - \dot{M}(\theta))$$

$$= (I + o(1))(\log n)^{-1/2}(P'QP)^{-1/2}(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n}(X_k-\theta)\varepsilon_k' + o(1). \quad (5.9)$$

We will need

$$(\log n)^{-1/2}\ \mathrm{vec}\Big((P'QP)^{-1/2}(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n}(X_k-\theta)\varepsilon_k'\Big)$$

$$\to_D N(0,(\Sigma \otimes I)). \quad (5.10)$$

To prove (5.10), we apply the Cramer-Wold device and Dvoretsky's (1972) result on the asymptotic normality of triangular martingale arrays. Let a_j be a $p\times 1$ column vector and $a' = (a_1',a_2',\ldots,a_p')$ where a is orthonormal. Then,

$$(\log n)^{-1/2}\ a'\mathrm{vec}\Big((P'QP)^{-1/2}(\dot{M}(\theta))^{1/2} \sum_{k=1}^{n}(X_k-\theta)\varepsilon_k'\Big) = \sum_{k=1}^{n} Z_{n,k}$$

where

$$Z_{n,k} = (\log n)^{-1/2} \sum_{j=1}^{p} a_j'(P'QP)^{-1/2}(\dot{M}(\theta))^{1/2}(X_k-\theta)\varepsilon_{k,j}.$$

Now, with $F_{n,k} = \sigma(\varepsilon_{\ell,j},\ j=1,\ldots,p,\ \ell=1,\ldots,k-1)$, we have $E(Z_{n,k}|F_{n,k-1}) = 0$ and

$$\sum_{k=1}^{n} E(Z_{n,k}^2|F_{n,k-1})$$

$$= \sum_{k=1}^{n}\sum_{r,s=1}^{p} a_r'(P'QP)^{-1/2}(\dot{M}(\theta))^{1/2}(X_k-\theta)$$

$$E(\varepsilon_{k,r}\,\varepsilon_{k,s}|F_{n,k-1})(X_k-\theta)'(\dot{M}(\theta))^{1/2}(P'QP)^{-1/2}a_s/(\log n)$$

$$\sim \sum_{r,s=1}^{p} a_r'(\Sigma)^{(rs)}a_s = a'(\Sigma \otimes I)a$$

by A5 and Lemma 4.5. With A5, the conditional Lindeberg condition is easy to check. This is sufficient to prove (5.10).

Thus, from (5.9) and (5.10),

$$(\log n)^{1/2}\ \mathrm{vec}\{(P'QP)^{1/2}(\dot{M}(\theta))^{-1/2}(\hat{\beta}_1-\dot{M}(\theta))\} \to_D N(0,(\Sigma \otimes I)).$$

which immediately implies the result. ‡

§6. Concluding Remarks

In stochastic approximation, as with other sequential designs, experimental design considerations are not necessarily identical to estimation considerations. Unlike previous authors, in this paper we have separated these two components. It is our hope that by showing the possible advantages of this separation we will open up a new area of investigation. Philosophically, our attitude is similar to Siegmund (1985) who emphasizes the different goals of stopping a classical sequential study early and estimation following the stopped study. Of course, in classical sequential analysis the observations are assumed to be i.i.d. while in stochastic approximation the design vector changes sequentially. Further, using a relatively simple design, we have shown how to efficiently estimate the root of an unknown function in a simple, heuristically appealing fashion. This separation to improve estimation procedures of the regression function maximum will be addressed in a later paper.

In this paper, the least-squares estimate has been proposed and justified based on asymptotic considerations. For some applications, it is desirable to

modify the estimators to achieve better performance, in some sense, in finite samples. For example, since our approach is to estimate the hyperplane around the root, we may wish to leave out the first few, say m, observations. The justification is that the initial observations are farther away from the root and bias the estimate. Let Ω^* be a n×n matrix, the lower right-hand submatrix being a (n-m) × (n-m) identity matrix and zeroes elsewhere. Define the weighted matrix of regression coefficients,

$$\beta^* = (X'\Omega^*X)^{-1}X'\Omega^*Y$$

$$= (\beta_0^* \ \beta_1^*)'.$$

It is easy to show that the weighted estimator,

$$\theta^* = -(\beta_1^*)^{-1} \beta_0^*,$$

enjoys the same asymptotic properties as $\hat{\theta}$ for fixed m. Anbar (1978) suggests leaving out initial observations in his adaptive stochastic approximation procedure. Another possibility is to use a bounded-influence regression estimator, for example, the Krasker-Welsch (1982) estimator. Such estimators automatically down-weight any observation with a Y that appears outlying relative to the regression model. Points with outlying X values are down-weighted more severely. A further possibility is transformation of the Y observations, as proposed by Anbar (1973) and Abdelhamid (1973). The transformation achieving asymptotic efficiency depends upon the density of the errors ε. Fabian (1973) has shown that when the density f is a priori unknown, one can still achieve asymptotic efficiency by estimating f. However, Fabian's algorithm estimates f sequentially during the course of experimentation, and this greatly increases the complexity of the algorithm. It may be possible to achieve asymptotic efficiency by using the untransformed Y observations during the sequential design and then estimating f only once, at the end of experimenta-

tion when θ is estimated. Our point here is that the recent literature on robust and adaptive estimation provides a variety of tools for fitting a regression model to data. By not insisting upon X_n as the final estimate of θ, these tools are made available after stochastic approximation is used to generate the design.

The rate of convergence of the gradient estimator in Theorem 3.4 is relatively slow in comparison to the usual square root n rate that is predominant in statistical large sample theory. It can be shown by using a Venter (1967) type modification of the Robbins-Monro procedure, taking additional observations at each step, the rate can be improved. The rate depends on how far apart the observations at each stage are selected and is similar to Lai and Robbin's (1979) Theorem 6. This suggests using additional observations at each stage to estimate higher order derivatives of the regression function. Knowledge of higher order derivatives gives the researcher more information about the local behavior of the regression function around the root. We intend to explore this issue in a later paper.

References

Abdelhamid, S. N. (1973). Transformation of observations in stochastic approximation. Ann. Statist. 1, 1158-1174.

Anbar, D. (1973). On optimal estimation methods using stochastic approximation. Ann. Statist. 1, 1175-1184.

Anbar, D. (1978). A stochastic Newton-Raphson method. J. Statist. Plan. Inf. 2, 153-163.

Blum, J. (1954a). Multidimensional stochastic approximation methods. Ann. Math. Statist. 25, 737-744.

Blum, J. (1954b). Approximation methods which converge with probability one. Ann. Math. Statist. 25, 382-386.

Bojanic, R. and Seneta, E. (1973). A unified theory of regular varying sequences. Math. Zeitschrift 134, 91-106.

Chow, Y. (1965). Local convergences of martingales and the law of large numbers. Ann. Math. Statist. 36, 552-558.

Chung, K. (1954). On a stochastic approximation method. Ann. Math. Statist. 25, 463-483.

Davis, M. (1971). Comparison of sequential bioassays in small samples. J. Royal Statist. Soc. B 33, 78-87.

Dvoretsky, A. (1972). Central limit theorems for dependent random variables. Proc. Sixth Berkeley Symp. Math. Statist. Prob. (Ed. L. LeCam et al.). Los Angeles: University of California Press, Vol. 2, 513-555.

Fabian, V. (1968). On asymptotic normality in stochastic approximation. Ann. Math. Statist. 39, 1327-1332.

Fabian, V. (1973). Asymptotically efficient stochastic approximation; the R-M case. Ann. Statist. 1, 486-495.

Krasker, W. and Welsch, R. (1982). Efficient bounded-influence regression estimation. J. Amer. Statist. Assoc. 77, 595-604.

Lai, T. and Robbins, H. (1979). Adaptive designs and stochastic approximation. Ann. Statist. 7, 1196-1221.

Lai, T. and Robbins, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. Z. Wahrsch. verw. Gebiete 56, 329-366.

Ljung, L. (1978). Strong convergence of a stochastic approximation algorithm. Ann. Statist. 6, 680-696.

Ljung, L. and Söderström, T. (1983). Theory and Practice of Recursive Identification. MIT Press, Cambridge, MA.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. Ann. Math. Statist. 22, 400-407.

Ruppert, D. (1985). A Newton-Raphson version of the multivariate Robbins-Monro procedure. Ann. Statist. 13, 236-245.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. Ann. Math. Statist. 29, 373-405.

Siegmund, D. (1985). Sequential Analysis: Tests and Confidence Intervals. Springer-Verlag, New York.

Venter, J. (1967). An extension of the Robbins-Monro procedure. Ann. Math. Statist. 38, 181-190.

Wei, C. (1985). Asymptotic properties of least squares estimates in stochastic regression models. Ann. Statist. (to appear).

Wetherill, G. (1966). Sequential Methods in Statistics. Metheun, London.

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