ADAPTIVE ESTIMATION VIA MARTINGALES

by

Tze Leung Lai

TECHNICAL REPORT NO. 25
OCTOBER 1992

Prepared Under Grant MDA 904-89-H-2040
For The National Security Agency

Reproduction in Whole or in Part is Permitted
for any Purpose of the United States Government

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ADAPTIVE ESTIMATION VIA MARTINGALES

by

Tze Leung Lai

TECHNICAL REPORT NO. 25
OCTOBER 1992

PREPARED UNDER GRANT MDA 904-89-H-2040
FOR THE NATIONAL SECURITY AGENCY

Also supported by the Air Force Office of Scientific Research, and by National Science Foundation Grant DMS91-04432 and issued as Technical Report No. 6, Department of Statistics, Stanford University.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ADAPTIVE ESTIMATION VIA MARTINGALES

TZE LEUNG LAI

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

Abstract

Martingale representations and martingale limit theorems provide powerful and natural techniques in the construction and analysis of adaptive estimators. We show here how martingale theory provides a simple, unified treatment of (i) asymptotically efficient adaptive choice of score functions for rank tests and estimators in censored and/or truncated regression models, and (ii) adaptive estimation of time series parameters using efficient recursive algorithms.
1. INTRODUCTION

In the Third Purdue Symposium on Statistical Decision Theory and Related Topics, Lai and Robbins (1982a,b) gave asymptotically optimal solutions to the following multi-period control problem in the econometrics literature. Consider the linear regression model

\[ y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, \]  

(1)

where \( \alpha \) and \( \beta \neq 0 \) are unknown parameters and the errors \( \epsilon_1, \epsilon_2, \ldots \) are i.i.d. random variables with mean 0 and variance \( \sigma^2 \). The problem is to choose successive levels \( x_1, \ldots, x_n \) in (1) so that the outputs \( y_1, \ldots, y_n \) are as close as possible to a given target value \( y^* \), under the assumption that the \( \epsilon_i \) are normally distributed. Although one can in principle use dynamic programming to find the optimal rule if one puts a prior distribution on \( (\alpha, \beta) \), this approach has not been successful because of prohibitive computational and analytical difficulties. Without loss of generality, assume that \( y^* = 0 \) and rewrite (1) as \( y_i = \beta(x_i - \theta) + \epsilon_i \), where \( \theta = -\alpha/\beta \). If \( \theta \) were known the optimal rule would be to set \( x_i \equiv \theta \). Assuming that \( \theta \) is known to lie in an interval \([a, A]\), Anderson and Taylor (1976) proposed to estimate \( \theta \) at stage \( t \) by the maximum likelihood estimator

\[ \widehat{\theta}_t = \left\{ A \wedge (\bar{x}_t - \bar{y}_t/\beta_t) \right\} \vee a, \quad \widehat{\beta}_t = \frac{\sum_{i=1}^{t}(x_i - \bar{x}_t)y_i}{\sum_{i=1}^{t}(x_i - \bar{x}_t)^2}, \]  

(2)

and to replace the optimal rule \( x_i \equiv \theta \) that requires knowledge of \( \theta \) by its adaptive version \( x_{t+1} = \widehat{\theta}_t \).

Here and in the sequel, we use \( \wedge \) and \( \vee \) to denote minimum and maximum, respectively, and \( \bar{x}_t = \frac{1}{t} \sum_{i=1}^{t} x_i \). This adaptive rule would be "self-optimizing" (cf. Kumar and Varaiya, 1986) if \( \widehat{\theta}_t \to \theta \) a.s. However, as shown by Lai and Robbins (1982b), there is an event with positive probability on which \( \widehat{\theta}_t \) does not converge to \( \theta \) and on which \( \widehat{\beta}_t \) is consistently of the wrong sign. This led Lai and Robbins (1982a,b) to start by considering the case in which \( \beta \) is known so that the maximum likelihood estimator of \( \beta \) is \( \widehat{\beta}_t = \bar{x}_t - \bar{y}_t/\beta \). When \( \beta \) is unknown but upper and lower bounds with the same sign for \( \beta \) are known, they used a modified version \( b_t \) of the least squares estimate \( \widehat{\beta}_t \) and showed that the modified least squares rule

\[ x_{t+1} = \bar{x}_t - \bar{y}_t/b_t \]  

(3)

converges a.s. to \( \theta \) and has the same asymptotic properties as the rule

\[ x_{t+1} = \bar{x}_t - \bar{y}_t/\beta. \]  

(4)

\[ 2 \]
The analysis of the modified least squares rule (3) is quite difficult. There is, however, a much simpler way of adapting (4) which also has the advantage of not requiring known upper and lower bounds with the same sign for $\beta$ and which can also be easily generalized beyond the linear regression model (1). To begin with, rewrite (4) in the form of an equivalent recursion
\[
x_{t+1} = x_t - y_t/(t\beta),
\]
(5)
cf. Lemma 1 of Lai and Robbins (1979). Combining (5) with (1) gives
\[
x_{t+1} - \theta = (x_t - \theta)(1 - t^{-1}) - (t\beta)^{-1}\epsilon_t.
\]
(6)
When $\beta$ is unknown, replacing $\beta$ by an estimate $b_{t-1}$ which depends on the observations up to stage $t-1$ gives
\[
x_{t+1} = x_t - y_t/(tb_{t-1}).
\]
(7)
Again (7) can be expressed in the form $x_{t+1} - \theta = (x_t - \theta)(1 - \beta(b_{t-1})^{-1}) - (b_{t-1})^{-1}\epsilon_t$, whose analysis can be carried out by the same kind of martingale arguments commonly used to analyze stochastic approximation schemes of the form (6), noting that $\{(tb_{t-1})^{-1}\epsilon_t\}$ is still a martingale difference sequence (just like $\{(t\beta)^{-1}\epsilon_t\}$). This is the essence of the adaptive stochastic approximation theory of Lai and Robbins (1979), who showed that the stochastic approximation scheme (5) and its adaptive version (7) have the same asymptotic properties if $b_{t-1} \to \beta$ a.s., in the general regression model $y_i = M(x_i) + \epsilon_i$ with $M(\theta) = 0$ and $M'(\theta) = \beta \neq 0$.

While (4) and (5) are equivalent, their adaptive versions (3) and (7) are no longer equivalent, and it is much easier to adapt (5) because of its inherent martingale structure. The moral of this story can be applied to a wide variety of adaptive estimation and control problems. An important first step is a martingale-type representation analogous to the re-expression of (4) into (5) to which adaptation is applied. This step may also take the form of approximating the basic statistic involved by a martingale transform $\sum_{i=1}^{n} v_i(\beta)\epsilon_i$ or $\int_{-\infty}^{\infty} v(\beta, s) dM(s)$. Here, in the discrete-time case, $\{\epsilon_i\}$ is a martingale difference sequence with respect to some increasing sequence of $\sigma$-fields $\{\mathcal{F}_i\}$ and $v_i(\beta)$ is $\mathcal{F}_{i-1}$-measurable for any given value of a (possibly infinite-dimensional) parameter $\beta$ to which one has to adapt. In the continuous-time problem, $\{M(s)\}_{-\infty < s < \infty}$ is a martingale with respect to some filtration $\{\mathcal{F}(s)\}_{-\infty < s < \infty}$ and $\{v(\beta, s)\}_{-\infty < s < \infty}$ is a predictable process for every fixed $\beta$. Replacing $\beta$ by an $\mathcal{F}_{i-1}$-measurable estimate $\tilde{\beta}_{i-1}$ still gives a martingale $\sum_{i=1}^{n} v_i(\tilde{\beta}_{i-1})\epsilon_i$ in the discrete-time case, and a similar idea can also be used in the continuous-time case.

In Section 2 we show how this idea leads to relatively simple and yet asymptotically efficient adaptive choice of the score functions for rank statistics in regression analysis when the response
variables are not completely observable because of the presence of left truncation and right censorship. In Section 3 we review some recent work that applies this idea to adaptive estimation in time series models. These results reinforce the conclusion from our earlier work in adaptive stochastic approximation and the multi-period control problem that martingale techniques are very powerful and natural tools in the construction and analysis of adaptive estimators and adaptive control rules.

2. ADAPTIVE CHOICE OF SCORE FUNCTIONS FOR RANK TESTS AND ESTIMATORS IN CENSORED/TRUNCATED REGRESSION MODELS

Without assuming that the $\epsilon_i$ in (1) have mean 0, we shall let $\alpha = 0$ in (1). Let $e_i(b) = y_i - bx_i$ and let $R_i(b)$ be the rank of $e_i(b)$ in \{e_1(b), \ldots, e_n(b)\}. For a given score function $\phi : (0,1) \to \mathbb{R}$ satisfying $\int_0^1 \phi(u) du = 0$ and $\int_0^1 \phi^2(u) du < \infty$, the associated linear rank statistic is of the form

$$L_n,\phi(b) = \sum_{i=1}^n x_i \phi(R_i(b)/n) \text{ or } L_n,\phi(b) = \sum_{i=1}^n x_i \phi_n(R_i(b)/n), \tag{8}$$

where $\phi_n(j/n)$ is the expected value of $\phi$ evaluated at the $j$th order statistic in a sample of size $n$ from the uniform distribution on $(0,1)$. There is an extensive literature on the use of $L_n,\phi(0)$ as test statistics for testing the null hypothesis $\beta = 0$, and on estimation of $\beta$ based on the estimating equation $L_n,\phi(b) = 0$ (cf. Hájek and Šidák, 1967; Jurečková, 1969). From the mean and variance of the asymptotic normal distribution of $L_n,\phi(\beta + c/\sqrt{n})$, it follows that $\phi_F = -f' \circ F^{-1} / f \circ F^{-1}$ is an asymptotically optimal choice of the score function $\phi$, assuming that the distribution function $F$ of the $\epsilon_i$ has a differentiable density function $f$.

Since $F$, and therefore $\phi_F$ also, are typically unknown, it is natural to use the data to estimate $\phi_F$. This leads to the adaptive linear rank statistic $L_n,\hat{\phi}(b)$, where $\hat{\phi}$ is an estimate of $\phi_F$. However, the linear rank statistic (8) in which the nonrandom score function $\phi$ is replaced by a random $\hat{\phi}$ becomes exceedingly difficult to analyze unless $\hat{\phi}$ is independent of \{x_1, y_1, \ldots, x_n, y_n\}. For the problem of testing $H_0 : \beta = 0$, Hájek (1962) split a given sample of $N$ observations \{(x_i, y_i)\} into two subsamples, the first consisting of $n$ observations and the second consisting of $m = N - n$ observations such that $m \to \infty$ but $m = o(N)$ as $N \to \infty$, and used the second subsample to construct an estimate $\hat{\phi}$, which is then applied to the first subsample to construct $L_n,\hat{\phi}(0)$. He pointed out, however, that this approach is "of little use in practice because of slow convergence" due to the negligible additional sample size $m$ relative to $n$. For the problem of estimating $\beta$ in the two-sample case (i.e., $x_i = 0$ or 1) via the estimating equation $L_n,\hat{\phi}(b) = 0$, van Eeden (1970) and Beran (1974) extended Hájek's approach under some additional assumptions. In particular,
while van Eeden (1970) assumed \( \phi_F \) to be nondecreasing so that \( \hat{\phi} \) can also be chosen to be nondecreasing, Beran (1974) assumes the existence of a \( \sqrt{n} \)-consistent estimator so that \( L_{n,\phi}^{\wedge}(b) \) can be substituted by a much more tractable linear approximation in some \( O(n^{-1/2}) \)-neighborhood of the \( \sqrt{n} \)-consistent estimator. The resultant adaptive estimators are shown to have the same asymptotic normal distribution as that using the optimal score function \( \phi_F \).

Prentice (1978) generalized the linear rank statistics (8) to the regression setting in which the \( y_i \) in (1) are not completely observable due to right censorship. Instead of \( y_i \), one observes \( \tilde{y}_i \equiv y_i \land c_i \) and \( \delta_i \equiv I\{y_i \leq c_i\} \), where \( c_i \) are independent random variables such that \( \{\epsilon_i\} \) is independent of \( \{(x_i, c_i)\} \). Let \( e_i(b) = \tilde{y}_i - bx_i \) and let \( \hat{F}_{n,b} \) be the Kaplan-Meier curve based on \( (e_i(b), \delta_i)_{1 \leq i \leq n} \).

For the case of complete data, \( \hat{F}_{n,b} \) reduces to the empirical distribution function of \( (e_i(b))_{1 \leq i \leq n} \) and the term \( \phi(R_i(b)/n) \) in (8) can be expressed as \( \phi(\hat{F}_{n,b}(e_i(b))) \). For censored data, Prentice (1978) generalized (8) to the form

\[
L_{n,\phi}(b) = \sum_{i=1}^{n} x_i \delta_i \phi(\hat{F}, \delta_i (e_i(b))) + (1 - \delta_i) \Phi(\hat{F}, \delta_i (e_i(b))) \tag{9}
\]

where \( \Phi(u) = \int_{0}^{u} \phi(t) dt \), \( 0 \leq u < 1 \). He also conjectured that these statistics are asymptotically equivalent to the so-called “weighted log-rank statistics” of the form

\[
S_{n,\psi}(b) = \sum_{i=1}^{n} \delta_i \psi(\hat{F}, \delta_i (e_i(b))) \left\{ x_i - \frac{\sum_{j=1}^{n} x_j I(e_j(b) \geq e_i(b))}{\sum_{j=1}^{n} I(e_j(b) \geq e_i(b))} \right\} \tag{10}
\]

where \( \psi = \phi - \Phi \). The special case \( \psi \equiv 1 \) in (10) corresponds to Mantel’s (1966) log-rank statistic. Cuzick (1985) proved the validity of this conjecture. Under the assumptions that the \( x_i \) are bounded and that \( \phi \) is twice continuously differentiable on \((0, 1)\) such that \( t|\phi'(1-t)| + t^2|\phi''(1-t)| \leq K t^{-\delta} \) for some \( \delta < 1/2 \) and \( K > 0 \), he showed that \( |L_{n,\phi}(b) - S_{n,\psi}(b)| \leq c_n \) for some nonrandom constants \( c_n = o(\sqrt{n}) \).

Although it is very difficult to analyze the censored rank statistics (9) directly, their asymptotically equivalent approximations (10) are much easier to handle. In fact, \( S_{n,\psi}(\beta) \) can be expressed as a stochastic integral with respect to a continuous-time martingale and one can therefore use martingale central limit theorems to establish the asymptotic normality of \( S_{n,\psi}(\beta) \). Contiguity arguments can then be used to obtain the asymptotic normality of \( S_{n,\psi}(\beta + c/\sqrt{n}) \). More specifically, let \( \Lambda(t) = -\log(1 - F(t)) \) be the cumulative hazard function and let

\[
M_i(t) = I_{\{\epsilon_i \leq t, \delta_i = 1\}} - \int_{-\infty}^{t} I_{\{\epsilon_i \land (c_i - \beta x_i) \geq s\}} d\Lambda(s),
\]

\[
Y_n(t) = \sum_{i=1}^{n} I_{\{\epsilon_i \land (c_i - \beta x_i) \geq t\}}, \quad X_n(t) = \sum_{i=1}^{n} x_i I_{\{\epsilon_i \land (c_i - \beta x_i) \geq t\}}. \tag{11}
\]
Let $\xi_i = \epsilon_i \wedge (c_i - \beta x_i)$ and let $\mathcal{F}(t)$ be the complete $\sigma$-field generated by $z_i, \delta_i I_{\xi_i \leq s}, \xi_i I_{\xi_i \leq s}$, $1 \leq i \leq n$. Then $\{M_i(t), \mathcal{F}(t), -\infty < t < \infty\}$ is a martingale and $S_n, \psi(\beta) = T_n(\infty)$, where

$$T_n(t) = \sum_{i=1}^{n} \int_{-\infty}^{t} \psi(\hat{F}_n, \beta(s)) \left\{ z_i - \frac{X_n(s)}{Y_n(s)} \right\} I_{\{Y_n(s) > 0\}} dM_i(s).$$

Note that $X_n(s)$ and $Y_n(s)$ are left continuous in $s$. If $\psi$ is continuous and $\hat{F}_n, \beta$ is chosen to be the left continuous version of the Kaplan-Meier curve, then $\psi \circ \hat{F}_n, \beta$ is left continuous and $\{T_n(t), \mathcal{F}(t), -\infty < t < \infty\}$ is a martingale. Hence martingale theory and in particular martingale central limit theorems can be applied to analyze $S_n, \psi(\beta) = T_n(\infty)$, cf. Fleming and Harrington (1991). More generally, if $\psi_n(\cdot)$ is a predictable process with respect to the filtration $\{\mathcal{F}(t)\}_{-\infty < t < \infty}$, then $\{\sum_{i=1}^{n} \int_{-\infty}^{t} \psi_n(s) \{z_i - X_n(s)/Y_n(s)\} I_{\{Y_n(s) > 0\}} dM_i(s), \mathcal{F}(t), -\infty < t < \infty\}$ is a martingale with predictable variation process

$$\sum_{i=1}^{n} \int_{-\infty}^{t} \psi_n(s) \{z_i - X_n(s)/Y_n(s)\}^2 I_{\{Y_n(s) > 0, \epsilon_i \wedge (c_i - \beta x_i) \geq s\}} d\Lambda(s),$$

cf. Fleming and Harrington (1991). We first consider the problem of testing $H_0 : \beta = 0$ to show that adaptation is relatively easy in this martingale framework. Let $\lambda = \Lambda' = f/(1 - F)$ be the hazard function. Note that

$$\phi_F(F(s)) - \int_{F(s)}^{1} \phi(t) dt/(1 - F(s)) = -(f'/f + f/(1 - F))(s) = -\lambda'(s)/\lambda(s).$$

An asymptotically optimal choice of $\psi$ in the weighted log-rank statistic (10) is $\phi_F = - (\lambda'/\lambda) \circ F^{-1}$, which is asymptotically equivalent to the choice $\phi = \phi_F$ in the censored linear rank statistic (9). Alternatively, the same asymptotic properties are preserved if we replace the weight function $\psi_F \circ \hat{F}_n, b$ in (10) simply by $-\lambda'/\lambda$. For this choice of weights, the weighted log-rank statistic reduces at $b = \beta$ to

$$\sum_{i=1}^{n} \frac{-\lambda'(\epsilon_i(\beta))}{\lambda(\epsilon_i(\beta))} \delta_i \left\{ z_i - \frac{X_n(\epsilon_i(\beta))}{Y_n(\epsilon_i(\beta))} \right\} = \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{-\lambda'(s)}{\lambda(s)} \left\{ z_i - \frac{X_n(s)}{Y_n(s)} \right\} I_{\{Y_n(s) > 0\}} dM_i(s). \quad (12)$$

In ignorance of $\omega = -\lambda'/\lambda$, we can replace it in (12) by an estimate $\hat{\omega}$. This adaptively weighted log-rank statistic can be analyzed by the same martingale arguments as before, as long as $\hat{\omega}(\cdot)$ is a predictable process. Note that unlike Hájek's (1962) treatment of the adaptive linear rank statistic $L_n, \hat{\phi}(\beta)$ by requiring the estimate $\hat{\phi}$ to be based on an additional subsample that is independent of $\{x_1, y_1, \cdots, x_n, y_n\}$, we can base the estimate $\hat{\omega}$ here on the same sample $\{x_1, y_1, \delta_1, \cdots, x_n, y_n, \epsilon_n\}$ and no longer need to take an additional independent sample merely for the sake of estimating $\omega$. 

6
We now discuss how estimators $\hat{\omega}$ can be constructed to satisfy the predictability condition. Under the null hypothesis $H_0 : \beta = 0$, $y_i = \epsilon_i$ are i.i.d. with common hazard function $\lambda$. Kernel estimators of $\lambda$ and its derivatives based on the censored data $(\bar{y}_i, \delta_i) i = 1, \ldots, n$, are of the form
\begin{equation}
\hat{\lambda}^{(\nu)}(t) = b^{-(\nu+1)} \sum_{i=1}^{n} \delta(i) K_{\nu}(b^{-1}(t - \bar{y}(i)))/(n - i + 1),
\end{equation}
where $\bar{y}(1) \leq \cdots \leq \bar{y}(n)$ are the order statistics of the $\bar{y}_i$ and $\lambda^{(\nu)}$ denotes the $\nu$th derivative of the hazard function $\lambda$ ($\lambda^{(0)} = \lambda$). Consistency and other asymptotic properties of $\hat{\lambda}^{(\nu)}$ under certain regularity conditions were established by Ramsay-Hansen (1983), Tanner and Wong (1983), Yandell (1983) and Müller and Wang (1990a). In particular, Müller and Wang (1990a) considered kernels with support $[-1, 1]$ and such that $K_{\nu}$ belongs to the class $M_{\nu, \nu + h}(-1, 1)$ of functions of bounded variation (BV), where
\begin{equation}
M_{\nu,h}(a,b) = \{ f \in BV[a,b] : \int_{a}^{b} f(x)x^j dx = 0 \text{ if } 0 \leq j < k \text{ and } j \neq \nu, \int_{a}^{b} f(x)x^\nu dx = (-1)^\nu \nu!, \int_{a}^{b} f(x)x^k dx \neq 0 \}.
\end{equation}
They assumed the bandwidth $b = b_n$ to satisfy
\begin{equation}
b \to 0, nb^{2\nu + 1} \to \infty \text{ and } nb/(\log n)^2 \to \infty \text{ as } n \to \infty,
\end{equation}
and also considered locally adaptive choice of such bandwidths.

In order that the process \{$\hat{\lambda}^{(\nu)}(t)$\} be predictable with respect to the filtration \{$\mathcal{F}(t)$\}, $\hat{\lambda}^{(\nu)}(t)$ has to depend only on the observations $(\bar{y}_i, \delta_i)$ with $\bar{y}_i < t$. This can be accomplished by choosing the kernel $K_{\nu}$ in (13) such that $K_{\nu}(x) = 0$ if $x \leq 0$. Such one-sided kernels have been considered by Müller and Wang (1990b) in the context of estimating $\lambda^{(\nu)}$ when there is a single jump discontinuity in $\lambda^{(\nu)}$. In particular, they made use of orthonormal ultraspherical polynomials of order $r$ on $[-1, 1]$ to construct such one-sided kernels having $r(\geq 1)$ continuous derivatives on $\mathbb{R}$, with $r + \nu$ odd, and belonging to the class $M_{\nu, \nu + h}(0, 1)$. They also established the uniform strong consistency and other asymptotic properties of these one-sided kernel estimators $\hat{\lambda}^{(\nu)}(\cdot)$ that are predictable with respect to the filtration \{$\mathcal{F}(t)$\}$_{-\infty < t < \infty}$.

From these estimates $\hat{\lambda}^{(0)}$ and $\hat{\lambda}^{(1)}$ of $\lambda$ and $\lambda'$, we can estimate $\omega = -\lambda'/\lambda$ by
\begin{equation}
\hat{\omega}(t) = -\left(\hat{\lambda}^{(1)}(t)/\hat{\lambda}^{(0)}(t)\right)I_{\hat{\lambda}^{(0)}(t) \geq b^{1/\nu}, t \leq \bar{y}_n - b^{1/\nu}}.
\end{equation}
To test $H_0 : \beta = 0$, consider the adaptively weighted log-rank statistic
\begin{equation}
S_n = \sum_{i=1}^{n} \delta_i \bar{\omega}(\bar{y}_i) \left\{ x_i = \frac{\sum_{j=1}^{n} x_j I_{\{\bar{y}_j \geq \bar{y}_i\}}}{\sum_{j=1}^{n} I_{\{\bar{y}_j \geq \bar{y}_i\}}} \right\}.
\end{equation}
Under $H_0$, $n^{-1/2} S_n = T_n(\infty)$, where $T_n(t) = n^{-1/2} \sum_{i=1}^{n} \tilde{\mathcal{G}}(s) \{x_i - \bar{X}_n(s)/Y_n(s)\} I_{\{Y_n(s) > 0, c_i - \beta x_i \geq s\}} dM_i(s)$ is a martingale with predictable variation process

$$
n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{t} \tilde{\mathcal{G}}^2(s) \left\{x_i - \frac{X_n(s)}{Y_n(s)}\right\}^2 I_{\{Y_n(s) > 0, c_i - \beta x_i \geq s\}} d\Lambda(s). \tag{17}
$$

Under suitable regularity conditions on $f$, and assuming that the $(c_i, x_i)$ satisfy the condition

$$
\Gamma_r(s) = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} E(x_i^r I_{\{c_i - \beta x_i \geq s\}}) \text{ exists for } s < F^{-1}(1) \text{ and } r = 0, 1, 2, \tag{18}
$$

it can be shown that (17) converges in probability to

$$
v_t = \int_{-\infty}^{t} \omega^2(s) \{\Gamma_2(s) - \Gamma_1^2(s)/\Gamma_0(s)\} dF(s), \tag{19}
$$

cf. Lai and Ying (1991). Hence, by Rebolledo's (1980) martingale central limit theorem, $n^{-1/2} S_n = T_n(\infty)$ has a limiting normal distribution with mean 0 and variance $v_\infty$ under $H_0 : \beta = 0$. A contiguity argument similar to the proof of Proposition 5.3.1 of Gill (1980) can be used to show that under contiguous alternatives $H_n : \beta = c/\sqrt{n}$, $n^{-1/2} S_n$ has a limiting normal distribution with variance $v_\infty$ and mean $cv_\infty$. Note that the same arguments and results also hold for the asymptotically optimal weighted log-rank statistic (12), which corresponds to replacing $\tilde{\mathcal{G}}$ in (16) by $\omega$.

In the case of complete data (for which $c_i \equiv \infty$), $\Gamma_2(s) - \Gamma_1^2(s)/\Gamma_0(s) \equiv \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$ and

$$
v_\infty = \left\{ \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right\} \int_{-\infty}^{\infty} (f'/f)^2 dF,
$$

which agrees with the fact that (12) is asymptotically equivalent to $L_{n, \phi_F}(\beta)$, as shown by Cuzick (1985). The preceding discussion shows that the “martingale approximation” (12) of the optimal linear rank statistic $L_{n, \phi_F}(\beta)$ greatly facilitates adaptation when $\phi_F$ is unknown. Furthermore, for censored data, even if $\phi_F$ is assumed to be known, the censored rank statistic $L_{n, \phi_F}(\beta)$ defined in (9) is very difficult to analyze directly, and its martingale approximation (12) makes the problem tractable.

Lai and Ying (1992a) studied rank tests based on censored regression data and again arrived at a martingale-type representation similar to (12) for the rank statistics. Suppose that in the regression model, $(x_i, y_i)$ is observable only when $y_i \geq t_i$, where the truncation variables $t_i$ are such that $(x_i, t_i)$ are independent random vectors that are independent of $\{\epsilon_i\}$. Thus, the observed data consist of $(x_i^0, y_i^0, t_i^0)$ with $y_i^0 \geq t_i^0$. Extending the arguments of Prentice (1978) and Cuzick (1985)
for censored data, Lai and Ying (1992a) obtained the following analogue of the rank statistic (10) for truncated data:

\[
S_{n, \psi}(b) = \sum_{i=1}^{n} \psi(\hat{F}_{n,b}(e_{i}(b))) \left\{ x_{i}^{\circ} - \frac{\sum_{j=1}^{n} x_{j}^{\circ} I\{ e_{j}(b) > \tau_{i}(b) \geq \tau_{j}(b) \}}{\sum_{j=1}^{n} I\{ e_{j}(b) > \tau_{i}(b) \geq \tau_{j}(b) \}} \right\},
\]

(20)

where \( e_{i}(b) = y_{i}^{\circ} - bx_{i}^{\circ}, \ t_{i}^{\circ}(b) = t_{i}^{\circ} - bx_{i}^{\circ}, \ \psi = \phi - \Phi \) for a given score function \( \phi \) on \((0,1)\), and \( \Phi(u) = \int_{0}^{u} \phi(t) \text{d}t / (1 - u) \) as before. The \( \hat{F}_{n,b} \) in (20) is a product-limit estimator for truncated data and is analogous to the Kaplan-Meier estimator for censored data. Lai and Ying (1992a) showed that the martingale arguments to analyze (10) can be extended to analyze (20) at \( b = \beta \), and by contiguity also at \( b = \beta + c/\sqrt{n} \). Again, an asymptotically optimal choice of the score function \( \phi \) is \( \phi = \phi_{F} \) and this asymptotically optimal choice is also equivalent to replacing \( \psi \circ \hat{F}_{n,b} \) in (20) simply by \( \omega = -\lambda'/\lambda \). The \( \sigma \)-field \( \mathcal{F}(s) \) to be used in connection with this martingale representation is the complete \( \sigma \)-field generated by \( x_{i}, t_{i}, I\{ t_{i} - s \leq e_{i} < s \}, I\{ t_{i} - s \leq e_{i} \leq u \} \) \((u \leq s, i \geq 1)\). When \( \omega = -\lambda'/\lambda \) is unknown, we can construct a consistent estimate \( \hat{\omega} \) such that \( \{ \hat{\omega}(t) \} \) is predictable with respect to the filtration \( \{ \mathcal{F}(t) \} \) by a straightforward modification of the preceding arguments in the censored case.

The above discussion on adaptive choice of score functions in rank tests of \( H_{0} : \beta = 0 \) shows the relative ease in adaptation by using martingale representations of rank statistics. A closely related problem, studied by Lai and Ying (1991, 1992b), is asymptotically efficient estimation of the regression parameter \( \beta \) via adaptive determination of the asymptotically optimal weight function \( \omega(= -\lambda'/\lambda) \) in the estimating equation \( S_{n,\omega}(b) = 0 \). Lai and Ying (1991, 1992b) gave a unified treatment of (10) and (20) by considering the more general setting in which the response variables \( y_{i} \) in the multiple regression model \( y_{i} = \beta^T x_{i} + \epsilon_{i} \) are subject to both right censorship and left truncation. Here the observed data consist of \( (x_{i}^{\circ}, \delta_{i}^{\circ}, \tau_{i}^{\circ}, e_{i}^{\circ}, I\{ \tau_{i}^{\circ} \leq e_{i}^{\circ} < \tau_{i}^{\circ} \}, I\{ \tau_{i}^{\circ} \leq e_{i}^{\circ} \leq \tau_{i}^{\circ} \}) \) with \( \delta_{i}^{\circ} \geq \tau_{i}^{\circ} \). Let \( e_{i}(b) = y_{i}^{\circ} - bx_{i}^{\circ} \). Estimation of the hazard function \( \lambda \) and its derivative \( \lambda' \) is more difficult in the present setting than in the preceding discussion on adaptive rank tests of \( H_{0} : \beta = 0 \). In the preceding discussion, we used the fact that under \( H_{0} : \beta = 0 \), the \( \epsilon_{i}(= y_{i}) \) are observable if they are neither censored nor truncated. For the problem of estimating \( \beta \), even if \( x_{i} \) and \( y_{i} \) are completely observable, \( \epsilon_{i} = y_{i} - \beta^T x_{i} \) is unobservable since \( \beta \) is unknown. Replacing \( \epsilon_{i} \) by \( e_{i}(b) \) and using the \( e_{i}(b) \) to estimate \( \omega \) leads to an intractable estimating equation of the form \( S_{n,\omega}(b) = 0 \), since the estimate \( \hat{\omega}_{b} \) of \( \omega \) depends on \( b \) through the \( e_{i}(b) \). An alternative approach is to estimate \( \omega \) from \( (\hat{\delta}_{i}^{\circ} - \hat{\beta}^T \hat{x}_{i}, \hat{\tau}_{i}^{\circ} - \hat{\beta}^T \hat{x}_{i})_{1 \leq i \leq n} \), where \( \hat{\beta} \) is some preliminary consistent estimator of \( \beta \). However, the \( \hat{\omega}(\cdot) \) thus constructed no longer satisfies the predictability condition since it involves \( \hat{\beta} \) which depends on all \( n \) observations.
To circumvent these difficulties caused by the need of a preliminary estimator of $\beta$ in the estimation of $\omega$, Lai and Ying (1991) used the following idea, assuming that $\omega$ is continuously differentiable. Divide the sample into two disjoint subsets, the first consisting of $n_1 = [n/2]$ observations and the second consisting of the remaining $n_2 = n - n_1$ observations. On the basis of only the observations of the first subsample, construct an estimate $\hat{\omega}_1$ of $\omega$ and apply it to the data in the second subsample to form the rank statistic $S^I_{n_2, \hat{\omega}_1}(b)$. Likewise, construct an estimate $\hat{\omega}_2$ of $\omega$ on the basis of the second subsample alone and apply it to the data in the first subsample to form the rank statistic $S^I_{n_1, \hat{\omega}_2}(b)$. Then combine these statistics based on the two separate subsamples into a single rank statistic
\[ S_n(b) = S^I_{n_1, \hat{\omega}_2}(b) + S^I_{n_2, \hat{\omega}_1}(b). \] (21)

Again martingale theory can be applied as before to analyze $n^{-1/2}S_n(\beta)$. Moreover, by using empirical process theory together with certain smoothing kernels in the definition of $\hat{\omega}_1$, $\hat{\omega}_2$, $S^I_{n_1, \hat{\omega}_2}(b)$ and $S^I_{n_2, \hat{\omega}_1}(b)$, Lai and Ying (1991) were also able to establish the asymptotic linearity of $S_n(b)$ as $n \to \infty$ and $b \to \beta$. Combining this asymptotic linearity property with the limiting normal distribution of $n^{-1/2}S_n(\beta)$ then shows that the estimating equation $S_n(b) = 0$ defines an asymptotically normal estimator $\hat{\beta}_n$ of $\beta$. Furthermore, Lai and Ying (1992b) showed that $\hat{\beta}_n$ is asymptotically efficient in the sense of attaining the asymptotic minimax bound for the present semiparametric estimation problem and also in the sense that the limiting distribution of $\sqrt{n}(T_n - \beta)$ is the convolution of some distribution with the limiting normal distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$, for any regular estimator $T_n$ of $\beta$.

3. ADAPTIVE ESTIMATION IN ARMAX MODELS

The subject of adaptive estimation began with Stein's (1956) work on estimating and testing hypotheses about a Euclidean parameter $\theta$ or, more generally, a function $h(\theta)$ in the presence of an infinite-dimensional nuisance parameter $G$. He considered the problem concerning when and how one can estimate $\theta$ when $G$ is unknown as well asymptotically as when $G$ is known. In the case of i.i.d. observations, Bickel (1982) and Fabian and Hannan (1982) provided general solutions to this problem. Their theory was subsequently extended by Kreiss (1987) to study adaptive estimation of the parameters of a stationary ARMA model. We shall consider the somewhat more general case of an ARMAX model (autoregressive moving average model with exogenous inputs) defined by the linear stochastic difference equation
\[ A(q^{-1})y_n = q^{-d}B(q^{-1})u_n + C(q^{-1})\epsilon_n, \] (22)
where \( \{u_n\}, \{y_n\} \) and \( \{\epsilon_n\} \) denote the input, output and disturbance sequences, respectively, \( d \geq 1 \) represents the delay, and
\[ A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_p q^{-p}, \quad B(q^{-1}) = b_1 + \cdots + b_k q^{-(k-1)}, \]
\[ C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h} \]
are scalar polynomials in the backward shift operator \( q^{-1} \)
(defined by \( q^{-1} x_n = x_{n-1} \)). The special case \( B(q^{-1}) = 0 \) reduces to the ARMA model, while the case \( C(q^{-1}) = 1 \) corresponds to the ARX model. Throughout the sequel the \( \epsilon_i \) are assumed to be i.i.d. symmetric random variables with a common continuously differentiable density function \( f \) such that
\[ I(f) = \int_{-\infty}^{\infty} \left( f'(x)^2 / f(x) \right)^2 f(x) \, dx < \infty. \]

Let \( \theta = (-a_1, \ldots, -a_p, b_1, \ldots, b_k, c_1, \ldots, c_h)^T \) and let
\[ \psi_i(\theta) = (y_{i-1}, \ldots, y_{i-p}, u_{i-d}, \ldots, u_{i-d-k+1}, \epsilon_{i-1}\theta, \ldots, \epsilon_{i-h}(\theta))^T, \]
where the \( \epsilon_i(\theta) \) are determined inductively by \( \theta \) and the \( y_i, u_i \) via the equation
\[ C(q^{-1}) \epsilon_i(\theta) = A(q^{-1}) y_i - B(q^{-1}) u_{i-d}. \]
Suppose that the input \( u_n \) has conditional density function \( p_n(y_n | u_{n-1}, \ldots, y_1, u_1) \) given the current and past data \( y_n, u_{n-1}, \ldots, y_1, u_1 \) and the initial condition \( x_0 = (y_0, \ldots, y_{1-r}, u_0, \ldots, u_{2-d-k}, \epsilon_0, \ldots, \epsilon_{1-h}) \) and that the \( p_n(n \geq 0) \) do not involve the parameter vector \( \theta \), where \( p_n \) is the density function of \( x_0 \). The log-likelihood function \( \ell_n(\theta) \) at stage \( n \) is
\[ \ell_n(\theta) = \sum_{i=1}^{n} \log f(y_i - \theta^T \psi_i(\theta)) + \sum_{i=0}^{n-d} \log p_i(u_i | y_i, u_{i-1}, \ldots, y_1, u_1, x_0). \]

When \( f \) is known, letting \( g = -f'/f \), the maximum likelihood estimator of \( \theta \) solves the estimating equation
\[ \sum_{i=1}^{n} g(y_i - \theta^T \psi_i(\theta)) \nabla(\theta^T \psi_i(\theta)) = 0, \quad (23) \]
where the gradient vectors \( \nabla(\theta^T \psi_i(\theta)) \) can be determined recursively by
\[ \nabla(\theta^T \psi_i(\theta)) + c_1 \nabla(\theta^T \psi_{i-1}(\theta)) + \cdots + c_h \nabla(\theta^T \psi_{i-h}(\theta)) = \psi_i(\theta). \quad (24) \]

For the ARMA model (i.e., \( B(q^{-1}) = 0 \)), Kreiss (1987) assumes that the zeros of \( A(z) \) and of \( B(z) \) all lie outside the unit circle, so that for some \( \eta > 1 \),
\[ 1/A(z) = \sum_{j=0}^{\infty} \alpha_j(\theta) z^j \quad \text{and} \quad 1/C(z) = \sum_{j=0}^{\infty} \beta_j(\theta) z^j \quad \text{for all} \quad |z| \leq \eta. \quad (25) \]

With \( \beta_j(\theta) \) defined from the invertibility of \( C(z) \), let
\[ Z_{t-1}(\theta) = \sum_{i=0}^{t-1} \beta_i(\theta) \psi_{t-i}(\theta), \quad (26) \]
which plays a basic role in Kreiss' construction of adaptive estimators. Under the identifiability assumption that \( a_p \neq 0, \ c_h \neq 0 \) and \( A(z), C(z) \) have no common zero and under some additional
assumptions, Kreiss (1987) showed that when $f$ is unknown, it is possible to construct adaptive estimators which have the same asymptotic optimality properties as the maximum likelihood estimator that assumes $f$ to be known and that is constrained to lie in a $O(n^{-1/2})$-neighborhood of the true parameter $\theta$. Starting with a $\sqrt{n}$-consistent preliminary estimator $\hat{\theta}_n$ of $\theta$ such that $\hat{\theta}_n$ can assume values in a prescribed discrete set, he defined an adaptive estimator $\hat{\theta}_n$ by a linear approximation around $\hat{\theta}_n$ to the estimating equation (23) in which the unknown $g$ is replaced by an estimate $\hat{g}_{n,i}$:

$$
\hat{\theta}_n = \hat{\theta}_n + \left( \hat{I}_n \sum_{i=1}^{n} Z_{i-1}(\hat{\theta}_n)Z_{i-1}^T(\hat{\theta}_n) \right)^{-1} \sum_{i=1}^{n} \hat{g}_{n,i}(\epsilon_i(\hat{\theta}_n))Z_{i-1}^T(\hat{\theta}_n),
$$  

(27)

where $\hat{I}_n$ is an estimate of $I(f) = \int (f'/f)^2 fdz$. The estimator $\hat{g}_{n,i}(x)$ of $g(x)$ in (27) depends on $x$ and the $n-1$ quantities $\epsilon_1(\hat{\theta}_n), \ldots, \epsilon_{i-1}(\hat{\theta}_n), \epsilon_{i+1}(\hat{\theta}_n), \ldots, \epsilon_n(\hat{\theta}_n)$.

The adaptive estimator (27) uses ideas similar to those of Bickel (1982) and Fabian and Hannan (1982) in the i.i.d. case, but is much more complicated because it involves the inversion (25) of the operator $C(q^{-1})$ for evaluating the $Z_{i-1}(\hat{\theta}_n)$ in (27). This is particularly inconvenient when one needs to update the estimators $\hat{\theta}_n$ sequentially whenever new data become available. The heavy computational burden in the estimator (27) is in sharp contrast to the recursive "on-line" estimators emphasized in the engineering literature, where the primary purpose of parameter estimation for the ARMAX system (22) is to support decisions that have to be taken on-line, i.e., during the operation of the system.

A large number of recursive estimation algorithms have been developed in the engineering literature. In particular, for the maximum likelihood equation (23) in the case of Gaussian $f$ with unit variance (for which $g(x) = -(f'/f)(x) = x$), Åström and Söderström proposed the following recursive approximation to the maximum likelihood estimator:

$$
\hat{\theta}_n = \theta_{n-1} + P_n \xi_n e_n,
$$  

(28)

where $\theta_n = (a_{n,1}, \ldots, a_{n,p}, b_{n,1}, \ldots, b_{n,k}, c_{n,1}, \ldots, c_{n,h})^T$ and

$$
e_n = y_n - \theta_{n-1}^T \phi_n, \quad \phi_n = (y_{n-1}, \ldots, y_{n-p}, u_{n-d}, \ldots, u_{n-d-k+1}, e_{n-1}, \ldots, e_{n-h})^T,
$$

$$
\xi_n + c_{n-1,1} \xi_{n-1} + \cdots + c_{n-1,h} \xi_{n-h} = \phi_n,
$$

$$
P_n = P_{n-1} - P_{n-1} \xi_n \xi_n^T / (1 + \xi_n^T P_{n-1} \xi_n),
$$

(cf. Ljung and Söderström (1983). No consistency or asymptotic normality results have been established for this algorithm, often called the RML2 (recursive maximum likelihood of the second
type) algorithm, because of analytical difficulties. Recently Lai and Ying (1992c) suggested using an auxiliary consistent estimate \( \tilde{\theta}_n (= \theta + o(\delta_n) \, \text{a.s.}) \) to monitor the RML2 algorithm, choosing \( \delta_n \) to be random variables that depend on the data up to stage \( n \) and such that \( \delta_n \to 0 \) and \( (n/\log n)^{1/2} \delta_n \to \infty \, \text{a.s.} \). They also showed how recursive estimators \( \tilde{\theta}_n \) can be constructed by the method of moments to satisfy these consistency conditions under certain assumptions. The monitored RML2 algorithm

\[
\begin{align*}
\theta_n &= \theta_{n-1} + P_n \xi_n e_n \quad \text{if} \quad \|\theta_{n-1} + P_n \xi_n e_n - \tilde{\theta}_n\| \leq \delta_n, \\
&= \tilde{\theta}_n \quad \text{otherwise},
\end{align*}
\]

becomes much more tractable, and is asymptotically normal and asymptotically efficient under certain regularity conditions, cf. Lai and Ying (1992c).

For the case of non-Gaussian but symmetric \( f \), let \( \omega(x) = g(x)/I(f) \), where \( g = -f'/f \) and \( I(f) = \int_{-\infty}^{\infty} g^2(x) f(x) dx \). A natural extension of the RML2 algorithm (28) to the present case is \( \theta_n = \theta_{n-1} + P_n \xi_n \omega(e_n) \), cf. Lai and Ying (1992c), where \( e_n, P_n \) and \( \xi_n \) are given in (29). Monitoring this by an auxiliary consistent estimator yields the following extension of (30):

\[
\begin{align*}
\theta_n &= \theta_{n-1} + P_n \xi_n \omega(e_n) \quad \text{if} \quad \|\theta_{n-1} + P_n \xi_n \omega(e_n) - \tilde{\theta}_n\| \leq \delta_n, \\
&= \tilde{\theta}_n \quad \text{otherwise}.
\end{align*}
\]

This is again asymptotically normal and asymptotically efficient under certain regularity conditions, as shown in Lai and Ying (1992c) by making use of martingale central limit theorems. The martingale structure which they use follows from the fact that \( \omega \) is an odd function and \( \epsilon_n \) is symmetric so that \( E\{P_n \xi_n \omega(e_n)|\mathcal{F}_{n-1}\} = P_n \xi_n E(\omega(\epsilon_n)) = 0 \), noting that \( \xi_n \) and \( P_n \) are \( \mathcal{F}_{n-1} \)-measurable by (29), where \( \mathcal{F}_n \) is the \( \sigma \)-field generated by \( x_0, u_1, y_1, \ldots, u_n, y_n \).

In ignorance of \( \omega \), if we replace it by an \( \mathcal{F}_{n-1} \)-measurable estimator \( \tilde{\omega}_{n-1} \) so that \( \tilde{\omega}_{n-1}(\cdot) \) is an odd function, then since \( \epsilon_n \) is symmetric and is independent of \( \mathcal{F}_{n-1} \), we again have

\[
E\{P_n \xi_n \tilde{\omega}_{n-1}(e_n)|\mathcal{F}_{n-1}\} = P_n \xi_n E\{\tilde{\omega}_{n-1}(e_n)|\mathcal{F}_{n-1}\} = 0.
\]

Hence we can again use the same martingale arguments as before. This enables Lai and Ying (1992c) to show that under certain assumptions the adaptive recursive estimator

\[
\begin{align*}
\theta_n &= \theta_{n-1} + P_n \xi_n \tilde{\omega}_{n-1}(e_n) \quad \text{if} \quad \|\theta_{n-1} + P_n \xi_n \tilde{\omega}_{n-1}(e_n) - \tilde{\theta}_n\| \leq \delta_n, \\
&= \tilde{\theta}_n \quad \text{otherwise},
\end{align*}
\]

has the same asymptotic properties as the recursive estimator (31) that assumes \( \omega \) to be known. Note that the estimator (32) is much simpler than Kreiss' adaptive estimator (27) in the ARMA
case. Furthermore, the $\hat{a}_{n-1}$ in (32) need only be updated periodically at times $m_1 < m_2 < \cdots$ to accommodate the more intensive computations involved, and there are also recursive algorithms for updating them, cf. Lai and Ying (1992c).

We can regard the recursive algorithm (31) as a martingale-type approximation to the estimating equation (23). Analogous to the relative ease in adapting the stochastic approximation scheme (5) instead of the equivalent recursion (4) to handle the case of unknown $\beta$, the recursive algorithms (31) and (32) again show the simplicity and elegance in using martingale representations/approximations to solve adaptive estimation problems.
REFERENCES


