Adaptive estimation in time series regression models

Douglas G. Steigerwald*

University of California, Santa Barbara, CA 93106, USA

This work develops adaptive estimators for a linear regression model with serially correlated errors. We show that these results continue to hold when the order of the ARMA process characterizing the errors is unknown. The finite sample results are promising, indicating that substantial efficiency gains may be possible for samples as small as 50 observations. We use these estimators to investigate the behavior of the forward foreign exchange market.

1. Introduction

The standard linear model has served as the workhorse of many economet-
ric studies. Within this framework most economists are familiar with the use
of the ordinary least squares (OLS) technique to estimate the slope parame-
ters. The popularity of this method stems from the fact that, when the errors
have a likelihood function that declines monotonically in the sum of squared
errors, OLS is equivalent to the method of maximum likelihood (ML). These
estimators are the efficient unbiased estimators; they have the smallest
possible asymptotic variance within the class of unbiased estimators.

Statisticians have long noted the efficiency losses of OLS when the errors
are nonnormal and, increasingly, empirical studies in economics have pointed
to the difficulty in assuming that the residuals have a Gaussian distribution.
An alternative approach involves the estimation of models with both para-
metric and nonparametric components. As a simple example consider the
linear model in which the conditional mean of the dependent variable,

denoted $X_t\beta$, is fully parameterized while the distribution of the error term is unspecified making this component nonparametric. Any estimator of β that can be formed without relying upon the unknown error density is a semiparametric estimator of β, one example of which is the OLS estimator.

In general, semiparametric estimators have larger asymptotic variances than the corresponding ML estimators based upon fully parameterized models, reflecting the loss of information inherent in leaving the error distribution unspecified. In certain cases however, the asymptotic distributions of these estimators are equivalent, and in this situation the semiparametric estimator is referred to as an adaptive estimator.

Adaptive estimators are especially attractive for use in empirical financial models. These situations are characterized by significant departures from normality and with the quantity of data available, precise estimation is an important goal. One could attempt to capture these departures by employing ML estimators with a specified distribution that is asymmetric or characterized by thick tails. Since the true distribution is not known these models are, in general, misspecified and their estimators are biased and inefficient. A more promising approach is to introduce conditional heteroscedasticity into the error sequence. In a recent survey article Bollerslev, Chou, and Kroner (1992) detail the extensive use of such models incorporating innovations with autoregressive conditional heteroscedasticity (ARCH). While an ARCH model depends upon only a small number of parameters to capture thick-tailed departures from normality, it does not account for asymmetry. Further, if this specification of the conditional moments is incorrect, the resultant misspecification is again a source of bias for the parameter estimators and leads to a variance that asymptotically exceeds the variance of the estimators of the correctly specified model. The semiparametric estimators introduced above rely upon much more general distributional assumptions, rendering them less susceptible to misspecification and restoring asymptotic efficiency in a wide range of cases.

Adaptive estimators have been developed for the linear model with independent errors in Bickel (1982) and for zero mean ARMA processes with known order in Kreiss (1987). Neither of these results are sufficient for use in empirical financial models. Such models typically have a nonzero conditional mean and include dependence in the error sequence, either through conditional second moments as in the ARCH design mentioned above, or through ARMA models with an unknown order. In this paper we prove the existence of adaptive estimators for general linear models in which the error process has an autoregressive moving average representation (ARMA) and explicitly consider both the case in which the order is known and in which it is unknown.

Section 2 begins by carefully defining the appropriate efficiency bound that an estimator must achieve to be adaptive. The class of models we consider is

described in section 3, while sections 4 and 5 show that the parameters of these models can be estimated adaptively. Finite sample results are presented in section 6, and in section 7 use the estimators to investigate the properties of the forward foreign exchange market. For the sake of both brevity and clarity, notational derivations and all proofs have been confined to the appendix.

2. Efficiency criterion

Adaptive estimators are semiparametric estimators that are asymptotically equivalent to ML estimators. Therefore a natural efficiency criterion would seem to be the Cramer–Rao lower bound. This is not tractable since superefficient estimators exist that have limiting distributions with a smaller variance than the MLE's at the point of superefficiency. For local neighborhoods of this point the relationship is reversed and the MLE has a smaller asymptotic variance [Hajek (1972)]. As a result we exclude superefficient estimators by restricting attention to estimators that have uniform limiting distributions in local neighborhoods of the true parameter value.

The corresponding efficiency criterion is based on the minimax principle. Consider the problem of estimating a parameter vector α using the estimator $\hat{\alpha}^T$ where T indexes the sample size. Given a loss function l, a sequence $\{\hat{\alpha}^T\}$ is locally asymptotically minimax if for any open interval around α the estimator approximately minimizes the maximum expected loss as the sample size becomes large.

More formally, to construct our minimax bound we let l represent a general loss function. This function maps $\mathbb{R}^a$ into $\mathbb{R}^+$, the set of nonnegative real numbers. We assume that $\{z: l(z) \leq c\}$ is closed, convex, and symmetric for every value of $c \geq 0$. We also require that

$$\int l(z)\phi_1(\lambda z)\, dz < \infty, \qquad \lambda > 0, \tag{2.1}$$

where ϕ_1 denotes the univariate standard normal density function.

We wish to examine local neighborhoods of α, so let A represent the set of all sequences which are $T^{1/2}$-convergent to α,

$$A = \{\{\alpha^T\}: |T^{1/2}(\alpha^T - \alpha) - m| \to 0, \forall m \in \mathbb{R}^a\}. \tag{2.2}$$

If our model is sufficiently regular that the log-likelihood function is quadratic over A, then it can be represented using a second-order Taylor series expansion and the estimation problem is locally asymptotically normal (LAN). Using a result from Fabian and Hannan (1982), when a problem is LAN its

asymptotic minimax bound is

$$\lim_{m \to \infty} \lim_{T \to \infty} \inf \sup_{|\alpha^T - \alpha| < m} E_{\alpha^T}\{l|T^{1/2}(\hat{\alpha}^T - \alpha^T)|\} \geq E\{l(Z)\}, \quad (2.3)$$

where $Z \sim N(0, \mathscr{I}(\alpha)^{-1})$ and $\mathscr{I}(\alpha)$ denotes Fisher's information matrix corresponding to the true distribution.

An estimator which achieves this bound asymptotically is a locally asymptotic minimax (LAM) estimator. An estimator is LAM if and only if its limiting distribution is invariant over A, giving us the following definition for an adaptive estimator.

Definition 2.1. A semiparametric estimator $\{\hat{\alpha}^T\}$ is adaptive if and only if

$$L_{\alpha^T}\{T^{1/2}(\hat{\alpha}^T - \alpha)\} \to N(0, \mathscr{I}(\alpha)^{-1}),$$

for all $\{\alpha^T\} \in A$, whenever $\mathscr{I}(\alpha)$ is finite and continuous at α.

3. Applications to regression problems

We will consider the linear regression model with an ARMA(p, q) error,

$$y_t = X_t\beta + \varepsilon_t, \quad (3.1)$$

where β is a subset of $\mathbb{R}^k$ and ε_t follows:

$$\varepsilon_t = u_t - \sum_{k=1}^{q} \theta_k u_{t-k} - \sum_{j=1}^{p} \rho_j \varepsilon_{t-j}. \quad (3.2)$$

Here $\{u_t\}$ is a sequence of independent identically distributed (i.i.d.) random variables with density function f. Let $\alpha = (\beta', \rho', \theta')'$ and note $\alpha \in \mathbb{R}^a$, where $a = k + p + q$.

To avoid nonparametric estimation of the density of the regressors, $\{X_t\}$, we must be sure that it is an ancillary statistic for the estimation of α. We therefore assume that weak exogeneity holds throughout so that we may confine our attention to the conditional density of $\{y_t\}$. In addition, we need to make the following assumptions:

Assumption 1. The time series error process $\{\varepsilon_t\}$ is stationary and invertible. The roots of the following polynomial equations are all strictly greater than 1 in modulus,

$$\theta(z) \equiv 1 + \theta_1 z + \cdots + \theta_q z^q = 0$$

and

$$\rho(z) \equiv 1 + \rho_1 z + \cdots + \rho_p z^p = 0.$$

Assumption 2. The independent and identically distributed random variables u_t have the common density f, with a zero mean and finite fourth moment. This density is absolutely continuous with respect to Lebesgue measure, $f(u) > 0$ for all u, f is symmetric, and $I(f) = \int (\dot{f})^2/f \, du$ is finite.

Assumption 3. For all $t = (1, 2, 3, \ldots)$ the common distribution $g_t(u_{1-p}, \ldots, u_0; \varepsilon_{1-p}, \ldots, \varepsilon_0; \varepsilon_1, \ldots, \varepsilon_T; \alpha)$ is absolutely continuous with respect to Lebesgue measure and is defined over the entire space on which α is defined. We also assume that the initial conditions are well behaved in that if $\hat{\alpha}^T$ is a $T^{1/2}$-consistent estimator of α, then

$$g_0(u_{1-q}, \ldots, u_0; \varepsilon_{1-p}, \ldots, \varepsilon_0; \hat{\alpha}^T) \to g_0(u_{1-q}, \ldots, u_0; \varepsilon_{1-p}, \ldots, \varepsilon_0; \alpha)$$

in P_α-probability.

Assumption 4. The $T \times k$ matrix of independent regressors, X, satisfies $T^{-1}\sum_{t=1}^{T} X_t' X_{t-s} \xrightarrow{\text{a.s.}} Q_s$, $s = 0, \ldots, p$, where Q_s is a nonsingular $k \times k$ positive definite matrix and $\xrightarrow{\text{a.s.}}$ indicates convergence almost surely. Note that X_t, the tth row of X, is an element of $\mathbb{R}^k$ corresponding to the observations on the independent variables for period t. While this specifically allows for serial correlation in our right-hand-side variables, we need this correlation to vanish asymptotically for the sequence $\{X_t\}$ to have a stationary time series representation. Therefore we require that the $\lim_{s \to \infty} Q_s = 0$ and $\sum_{s=0}^{\infty} Q_s = Q < \infty$. Since this last assumption restricts only the second moments of $\{X_t\}$, it is weaker than the restriction that $\{X_t\}$ be an α-mixing sequence.

Assumption 5. If we let $\xi_t = \dot{f}(u_t)/f(u_t)$, then ξ is assumed to satisfy

$$\lim_{\lambda \to 0} \int [\xi(u + \lambda) - \xi(u)] f(u) \, du = 0.$$

In addition, the expected value of the derivative of ξ should satisfy

$$\lim_{\lambda \to 0} \int [\xi(u - \lambda) - \xi(u)]/\lambda \, f(u) \, du = I(f).$$

We must first derive the asymptotic minimax bound for the model of eqs. (3.1) and (3.2). Using the logic outlined in section 2, if we can show that LAN holds, then our bound is given by (2.3). The LAN condition requires that our log-likelihood be asymptotically quadratic in a neighborhood of the true parameter value (if this property held over the entire parameter space, the model would be globally asymptotically normal). We define our neighbor-

hood, using the set A, as

$$\beta^T = \beta + T^{-1/2}m_1, \quad \rho^T = \rho + T^{-1/2}m_2, \quad \theta^T = \theta + T^{-1/2}m_3, \quad (3.3)$$

where $m_1 \in \mathbb{R}^k$, $m_2 \in \mathbb{R}^p$, and $m_3 \in \mathbb{R}^q$. Using these neighborhoods we can concisely write $u_t^T - u_t$ as

$$u_t^T - u_t = (m_1' : m_2' : m_3') \mathcal{S}_t(\alpha^T, \alpha), \tag{3.4}$$

where the derivation of S_t is described in the appendix.

For all points in the neighborhood A the log-likelihood ratio is

$$\ln L_T = \ln\left[\prod_{t=1}^{T} f^{1/2}(u_t^T) \Big/ f^{1/2}(u_t)\right],$$

with the score function

$$S_T(\alpha^T) = \sum_{t=1}^{T} \xi_t \mathcal{S}_t(\alpha^T). \tag{3.5}$$

Here $\mathcal{S}_t(\alpha^T)$ represents the derivative of u_t with respect to α evaluated at the point α^T, and we use $f^{1/2}$ rather than f because we want to work with square integrable functions and the definition of a density insures that $f^{1/2}$ is a member of this class.

Lemma 3.1.   Under Assumptions 1–5 the log-likelihood ratio for the linear model with ARMA(p, q) errors is such that

$$\ln L_T - \sum_{t=1}^{T} \xi(u_t)(u_t^T - u_t)$$

$$+ \tfrac{1}{2}(m_1' : m_2' : m_3') \mathcal{I}(\alpha)(m_1 : m_2 : m_3)' \to 0,$$

as $T \to \infty$.

A proof of Lemma 3.1 is given in Steigerwald (1989a). An important consequence of this lemma is

$$L_\alpha\{S_T(\alpha)\} \to N(0, \mathcal{I}(\alpha)). \tag{3.6}$$

4. Efficient estimation

We have now shown that one can find suitable conditions under which local asymptotic normality holds for the linear regression model with serially correlated disturbances. Under LAN, Fabian and Hannan (1982) have shown that a sequence of estimators $\{\hat{\alpha}^T\}$ is locally asymptotic minimax if

$$T^{1/2}(\hat{\alpha}^T - \alpha) - \mathscr{I}(\alpha)^{-1}S_T(\alpha) = o(1) \tag{4.1}$$

in probability under α.

In proving the convergence in (4.1) we will make use of the discretized estimator, $\bar{\alpha}^T$, developed by Le Cam. We use this to construct the discretized residuals which, upon setting $(\bar{u}_{T,0}, \ldots, \bar{u}_{T,1-q})$ and $(\bar{\varepsilon}_{T,0}, \ldots, \bar{\varepsilon}_{T,1-p})$ equal to zero, are obtained from

$$\bar{u}_{T,t} = \sum_{j=0}^{p} \bar{\rho}_{T,j}\bar{\varepsilon}_{T,t-j} + \sum_{k=1}^{q} \bar{\theta}_{T,k}\bar{u}_{T,t-k}, \tag{4.2}$$

where $\bar{\varepsilon}_{T,t} = y_t - X_t\bar{\beta}_T$. The empirical density, which assigns mass T^{-1} to each residual, is not continuous and poorly approximates f.

To overcome this we follow Bickel (1982) and convolute the empirical density with the density of a mean zero normal random variable, ζ_T. The variance of the normal variable, σ_T^2, is selected by the researcher to control the degree of smoothing. Since we will be examining a derivative of this density estimator, we need to define it over a small neighborhood around each of our residuals, $\bar{u}_{T,t}$. To do this let our smoothed density estimator be defined for all z in a small neighborhood of each $\bar{u}_{T,t}$ as

$$\hat{f}_{T,t}(z) = [2(T-1)]^{-1} \sum_{j=1, j \neq t}^{T} \left[\zeta_T(z + \bar{u}_{T,j}) + \zeta_T(z - \bar{u}_{T,j})\right]. \tag{4.3}$$

Now the derivative of the log of the density at u_t, $\xi(u_t, f) \equiv \xi_t$, can be estimated by

$$q_{T,t} = \xi(\bar{u}_{T,t}, \hat{f}_{T,t}) \quad \text{if} \quad |\bar{u}_{T,t}| < g_{1,T}, \quad \hat{f}_{T,t}(\bar{u}_{T,t}) \geq g_{2,T},$$

$$\text{and} \quad |\hat{f}'_{T,t}(\bar{u}_{T,t})| \leq g_{3,T} \, \hat{f}_{T,t}(\bar{u}_{T,t}),$$

$$q_{T,t} = 0 \qquad \text{otherwise,} \tag{4.4}$$

where our smoothing parameter, σ_T, and the trimming parameters $(g_{1,T}, g_{2,T}, g_{3,T})$ are all asymptotically negligible. To insure this we impose the following

condition:

Condition T.   The trimming parameters must satisfy $g_{1,T} \to \infty$, $g_{2,T} \to 0$, and $g_{3,T} \to \infty$ as $T \to \infty$. The smoothing parameter must satisfy $\sigma_T \to 0$ in such a way that $g_{3,T}\sigma_T \to 0$ and $g_{1,T}\sigma_T^{-3}$ is $o(T)$ as T tends to infinity.

To incorporate the kernel density estimator define $\hat{\alpha}^T$ as

$$\hat{\alpha}^T = \bar{\alpha}^T + \hat{\mathscr{I}}_T(\bar{\alpha}^T)^{-1}\hat{S}_T(\bar{\alpha}^T), \qquad (4.5)$$

which is analogous to the linearized likelihood estimator using estimators of the score function and information matrix. Our estimator of the score function is

$$\hat{S}_T(\bar{\alpha}^T) = T^{-1/2} \sum_{t=1}^{T} q_{T,t}(\bar{\alpha}^T)\mathscr{I}_t(\bar{\alpha}^T). \qquad (4.6)$$

To consistently estimate the matrix $\mathscr{I}(\bar{\alpha}^T)$ use $\hat{\mathscr{I}}_T(\bar{\alpha}^T) = \hat{I}_T(\bar{\alpha}^T)G_T(\bar{\alpha}^T)$, where

$$\hat{I}_T(\bar{\alpha}^T) = T^{-1} \sum_{t=1}^{T} q_{T,t}^2(\bar{\alpha}^T), \quad G_T(\bar{\alpha}^T) = \sum_{t=1}^{T} \mathscr{I}_t(\bar{\alpha}^T)\mathscr{I}_t(\bar{\alpha}^T)'. \quad (4.7)$$

The consistency of this estimator is shown in the proof of the following theorem:

Theorem 4.1.   Under Assumptions 1–5 and Condition T the estimator, $\{\hat{\alpha}^T\}$, for the linear regression model with ARMA(p, q) errors is such that $T^{1/2}(\hat{\alpha}^T - \alpha) - \mathscr{I}(\alpha)^{-1}S_T(\alpha)$ is $o(1)$.

   The above theorem insures that $\hat{\alpha}^T$ satisfies the convergence criterion of (4.1). It therefore is locally asymptotic minimax and, since it incorporates a nonparametric density estimator, is considered LAM-adaptive.
   The estimators defined in Theorem 4.1 are distinct from an estimator proposed by Stein (1960) for a related problem. For the linear regression model in which (y, X) have a joint multivariate normal distribution he showed that, when the number of independent regressors exceeds 2, the OLS estimators (which are MLE for this problem) are inadmissible when quadratic loss is the measure of risk. The biased estimator which dominates the MLE is a nonlinear function of the MLE and the multiple correlation coefficient for the problem. For the class of models we consider, (y, X) is typically non-Gaussian and the nonlinear adaptive estimator is equivalent to the MLE and is not dominated by the nonlinear biased estimators described above.

5. Adaptive estimation with unknown order

The theoretical results derived above required that the order of the ARMA process be known. While this assumption is satisfied in some empirical models, in most situations it cannot be maintained. Without it, one is naturally led to ask if it is still possible to construct adaptive estimators. Before we can answer this we need to carefully define the class of models we will consider.

We must exercise caution to restrict ourselves to identified models; it is this concern which precludes us from simply selecting an overparameterized model under the assumption that the excess parameters will have estimated values that are not statistically indistinguishable from zero. To understand this point, consider the parameter space corresponding to an ARMA$(1, 1)$ model. At first glance, the domain would seem to be the subset of $\mathbb{R}^2$ satisfying Assumption 1. Now consider points in this subspace along the line $\rho_1 = \theta_1$. For this set of points the transfer function $[\rho(L)^{-1}\theta(L)]$ is equal to 1, which is identical to a model in which the innovations are uncorrelated. If we tried to estimate ρ_1 and θ_1, our estimators would not converge to a unique point on the line $\rho_1 = \theta_1$.

To formally treat the problem of identification, the following notation will be helpful. A polynomial in the lag coefficients, $\phi(L)$, is a left divisor of $\rho(L)$ and $\theta(L)$ if

$$\rho(L) = \phi(L)\,\tilde{\rho}(L), \qquad \theta(L) = \phi(L)\,\bar{\theta}(L).$$

The lag polynomials $\rho(L)$ and $\theta(L)$ are said to be right multiples of $\phi(L)$, and $\phi(L)$ is the greatest common left divisor if it is a right multiple of all left divisors. To insure that our model has a unique representation we employ the following definition:

Definition 5.1.   For the ARMA (p, q) process described in (3.2), $\rho(L)$ and $\theta(L)$ have 1 as a greatest common left divisor.

It should be clear that this definition excludes models with common factors. We now modify Assumption 1 to restrict attention to identified models.

Assumption 1′.   The time series error process is stationary, invertible, and identified.

This assumption guarantees that the true model will be associated with unique values of p and q.

Given that our model is identified we must now confront two questions. First, is it possible to consistently estimate the order of the process? Second, if a consistent procedure is available, does the use of the procedure affect the asymptotic distribution of the estimators derived in section 4? We will address each of these questions in turn.

Methods that consistently estimate the order of the ARMA process have been developed in the statistics literature. We choose a stepwise Lagrange multiplier procedure, first proposed by Pötscher (1983), that uses the sum of the squared prediction errors as the criterion function. It is important that we use a stepwise testing procedure for, as Pötscher (1990) has noted, failure to do so can lead to a lack of identification under the alternative hypothesis. To see this, suppose we test the null hypothesis that the process is an ARMA(1, 1) against the alternative that it is an ARMA(2, 1). If we do not restrict our attention to identified models, then our null class includes models with $\rho_1 = \theta_1$. Observe that under H_0 when $\rho_1 = \theta_1$, both our restricted estimators and the Lagrange multiplier (LM) test statistic are inconsistent. Even if we confine our attention to null models that are identified, overparameterization under the alternative can create difficulties. Suppose we test $H_0 : \varepsilon_t$ is an ARMA(1, 0) against $H_1 : \varepsilon_t$ is an ARMA(1, 1). For almost every value of ρ_1 the rank of the covariance matrix used in constructing the LM test statistic is 1. However if $\rho_1 = 0$, the rank of this matrix is 0, and since there are no continuous generalized inverses over sets of matrices with different ranks, the LM test statistic is not consistent over the entire range of possibilities generated by H_0. Of course, if $\rho_1 = 0$, the appropriate test is $H_0 : \varepsilon_t$ is white noise versus $H_1 : \varepsilon_t$ is an ARMA(1, 0); our problem above stems from the fact that the white noise model is not identified under the ARMA(1, 1) alternative.

The testing strategy proceeds as follows. Select values (p^*, q^*) that are large enough to insure that the true model (p^0, q^0) is nested, hence $p^0 \le p^*$ and $q^0 \le q^*$. Construct a chain consisting of (p_i, q_i) where $p_1 = q_1 = 0$, and for each additional element (pair) in the chain either $p_{i+1} = p_i + 1$ and $q_{i+1} = q_i$ or $p_{i+1} = p_i$ and $q_{i+1} = q_i + 1$. For example,

$$\{(0,0),(1,0),(1,1),\ldots,(p^*,q^*)\}$$

is one such possible chain. At each step, beginning at $(0,0)$, estimate the ARMA models and construct an LM test of $H_0 : (p, q) = (p_i, q_i)$ vs. $H_1 : (p, q) = (p_{i+1}, q_{i+1})$. Choose the model which is associated with the first failure to reject. Do this for all possible chains and select the model with the smallest number of parameters (in case of ties, choose the model with the minimum value of q).

Given that it is possible to consistently estimate the order of the ARMA process, we turn our attention to the second question mentioned above. First,

observe that local asymptotic normality is a property that characterizes the underlying model whether or not p and q are known at the estimation stage.

Next we ask, does the linearized likelihood estimator, $\tilde{\alpha}^T$, satisfy eq. (4.1) when the order of the ARMA process characterizing ε is unknown? To answer this we call upon a result that is well known in the statistics literature, namely that consistent estimation of the order leaves the asymptotic distribution of the parameter estimators unchanged [Pötscher (1990)]. As a consequence we have:

Corollary 5.2. Under Assumptions 1'–5, the estimator $\tilde{\alpha}^T$ remains locally asymptotic minimax when the order of the ARMA process describing ε is unknown.

Finally, we must determine if it is still possible to construct adaptive estimators when the order is unknown. In the context of the above discussion, this will be true if the introduction of an unknown innovation density leaves the asymptotic equivalence argument discussed in the previous paragraph unchanged. This will follow if the procedure used to estimate the order cannot make use of any information associated with knowledge of the underlying distribution. Intuitively, the fact that the Lagrange multiplier test statistic is independent of $f(u)$ gives us our result. This is made more precise in the following theorem:

Theorem 5.3. For the linear regression model given in eqs. (2.1) and (2.2), Assumptions 1'–5 and Condition T insure that the estimator $\hat{\alpha}^T$ is adaptive when the order of the error process is unknown.

The use of a consistent testing procedure raises an interesting question. Given maximal values p^* and q^*, the corresponding process, an ARMA(p^*, q^*), is known as the overall model. The use of a consistent testing procedure leads to superefficient estimators for some elements of p^* and q^*. From our discussion in section 2 this might seem to violate the regularity conditions we imposed in deriving our efficiency bound. However, these estimators are superefficient only for the 'excess' parameters in the model, the parameters which always equal 0 when the order is known. The estimators are not superefficient for the parameters contained in α.

Further, it is important that we use a consistent procedure to insure that we have identifiability in the overall model. Suppose the true model is an ARMA(1, 0), identification of this model imposes the requirement that $\rho_1 \neq 0$. If our overall model is an ARMA(2, 1) and we let the parameters of this model be $(\delta_1, \delta_2, \gamma_1)$, then the transfer functions of the two models are identical if $\delta_1 - \gamma_1 = \rho_1$ and $\delta_2 - \rho_1\gamma_1 = 0$. In this case we need to select the true order of the model to guarantee that it is identified.

6. Simulations

All of the above results concern the limiting distributions of our estimators. Yet potential applied users need some guide to the performance of these estimators in finite samples. In an effort to provide this, the following Monte Carlo sampling experiment is used.

We consider the linear model with an intercept and one explanatory variable,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

Here x_t is a Bernoulli random variable with equal probability of assuming either 0 or 1 and is independent of ε_t. The error, ε_t, follows an AR(1) or an MA(1) with an innovation sequence, $\{u_t\}$, that is either Gaussian or drawn from one of the following densities (all of which are standardized to have mean 0 and variance 1). The first of these is a normal mixture that corresponds to the statistical definition of a contaminated sample in which some of the observations are, unknowingly to the researcher, characterized by additional sampling error. We draw 90% of the observations from a $N(0, \frac{1}{9})$ and 10% of the observations from a $N(0, 9)$ creating a density that is leptokurtic (it has a kurtosis that exceeds 3). It is similar to a Student-t distribution and differs from a normal random variable in two important ways: more of the density mass is concentrated near the mean and the tails are thicker. The next innovation distribution is a bimodal symmetric mixture, $0.5N(-3, 1) + 0.5N(3, 1)$. It corresponds to problems in which two distinct samples have been combined forming a density with a kurtosis that is less than 3 (it declines toward 1 as the distance between the center of the two underlying densities grows). This density also has thicker tails than the normal distribution, yet less of its density mass is concentrated near the mean. The final distribution is lognormal [$\exp(Z)$ where $Z \sim N(0, 1)$] which is both asymmetric and leptokurtic.

An experiment consists of generating 50 observations for both x and u and using this information to construct y. The generated data is then employed to estimate both the order of the model and the underlying parameters. The sequence is repeated 10,000 times for each possible experimental design. In constructing the adaptive estimators, we must specify both the smoothing and the trimming parameters. The smoothing parameter is selected via cross-validation using a loss function that corresponds to the normal kernel used in the nonparametric density estimator. The trimming parameters are chosen such that when the underlying density is normal, all three bind at the same value of u; $g_{1,T} = m$, $g_{2,T} = \exp(-m^2/2)$, and $g_{3,T} = m$. We choose $m = 8$ in accordance with Hsieh and Manski (1987).

Table 1

90% confidence intervals for $\hat{\beta}$.

	OLS	GLS	ADAPT	LLE
	$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$			
Normal	(0.47, 1.53)	(0.56, 1.45)	(0.48, 1.51)	(0.56, 1.44)
Unimodal	(0.47, 1.53)	(0.57, 1.45)	(0.63, 1.35)	(0.68, 1.36)
Bimodal	(0.48, 1.52)	(0.56, 1.42)	(0.70, 1.33)	(0.64, 1.34)
Lognormal	(0.48, 1.52)	(0.59, 1.42)	(0.60, 1.42)	(0.58, 1.45)
	$\varepsilon_t = u_t + 0.5u_{t-1}$			
Normal	(0.49, 1.51)	(0.53, 1.48)	(0.48, 1.53)	(0.55, 1.45)
Unimodal	(0.46, 1.53)	(0.59, 1.46)	(0.64, 1.36)	(0.65, 1.33)
Bimodal	(0.48, 1.52)	(0.52, 1.49)	(0.62, 1.39)	(0.61, 1.40)
Lognormal	(0.50, 1.52)	(0.57, 1.47)	(0.57, 1.43)	(0.56, 1.45)

The results of these experiments are reported in table 1, which lists the 90% confidence interval from the empirical distribution of $\hat{\beta}$. We focus our attention on the slope coefficient since it is most frequently of central concern to researchers. In judging the adaptive estimator it is helpful to make two comparisons. The first compares the confidence interval for the adaptive estimator with that of its GLS counterpart, revealing the efficiency gains arising from the use of this technique when the underlying density is nonnormal. The second comparison relates the efficiency gains of the adaptive estimator to those of the linearized likelihood estimator (LLE). The LLE is constructed using the one-step procedure of (4.5) with the actual values of the score function and information matrix, and this comparison points out how well the nonparametric estimator of the score function is performing.

Our results are broadly consistent across the two possible error processes, the AR(1) and MA(1). When the underlying innovation sequence is normal, the adaptive estimator has an efficiency loss of approximately 10% relative to the GLS estimator. Nonparametrically estimating the density when it is not necessary is comparable, in this example, to ignoring the serial dependence in the error process (the OLS and adaptive intervals are roughly equal).

Under the nonnormal distributions the results are quite promising. For the symmetric distributions, the adaptive estimators have confidence intervals between 20% and 40% smaller than their GLS counterparts based upon samples of only 50 observations. Further, the adaptive estimators capture most of the gains that are present as their confidence intervals closely approximate those of the LLE estimators. When the errors are lognormally distributed, the results are not as favorable. Since the earlier theorems hold only for symmetric densities, the construction of the nonparametric compo-

nent exploits this property in (4.3), and this reduces the efficiency of the adaptive estimator when the errors are not symmetric. The steep peak of this distribution also greatly increases q near the mode and the third trimming parameter removes many data points. This problem explains the relatively poor performance of the LLE in this experiment.

7. Forward rate unbiasedness

One of the interesting issues in international finance concerns the relationship between the forward rate and the spot rate for a specific currency. Economists have proposed that the efficient operating of this market should exclude arbitrage profit opportunities. As a result, deviations of the forward rate from the spot rate shoud be zero on average and unpredictable at the time the forward rate is set.

Numerous econometric studies have examined this relation empirically. While early studies defined the null hypothesis in terms of market efficiency, careful examination of the underlying models revealed a host of auxiliary assumptions. Recently, the null hypothesis has been more carefully defined in terms of a test that the forward rate is an unbiased predictor of the future spot rate.

To understand the relationship between the forward rate and the spot rate for a specific currency, consider the arbitrage pricing model developed by Cox, Ingersoll, and Ross (1981). For the following discussion, all of the variables are measured in dollars. Let $F_{t,q}$ denote the time t foreign currency price of a U.S. dollar delivered q periods in the future. If S_{t+q} represents the time $t+q$ foreign currency spot price for a U.S. dollar, then $S_{t+q} - F_{t,q}$ is the time $t+q$ payoff of $F_{t,q}$. Now let $R_{t,q}$ be the risk-free return on a q-period bill issued at time t. Cox, Ingersoll, and Ross have shown that $F_{t,q}$ is equal to the present discounted value of $S_{t+q}R_{t,q}$ using the following intuitive argument. Consider investing $F_{t,q}$ dollars in a riskless q-period bill at time t and $R_{t,q}$ dollars in q-period forward contracts. The initial investment is just $F_{t,q}$ dollars since the forward contracts are payable at time $t+q$. The resulting payoff at time $t+q$ is

$$F_{t,q}R_{t,q} + (S_{t+q} - F_{t,q})R_{t,q} = S_{t+q}R_{t,q}. \qquad (7.1)$$

Under arbitrage, an equilibrium is characterized by

$$F_{t,q} = E_t(S_{t+q}),$$

which states that the q-period forward rate must equal the time t value of a contract that pays $S_{t+q}R_{t,q}$ in period $t+q$. We let this expectation equal the

mathematical expectation defined over the time t information set available to the econometrician.

If we let f and s denote the natural logarithms of F and S, respectively, then $s_{t+q} - f_{t,q}$ represents the returns to holding foreign currency. Similarly, $s_t - f_{t,q}$, the deviation of the spot rate from the contemporaneous forward rate, captures all the information available to investors at time t. Under rational expectations this information should be orthogonal to the realized return leading to

$$s_{t+q} - f_{t,q} = \alpha + \beta(s_t - f_{t,q}) + \varepsilon_{t+q,q}. \tag{7.2}$$

Under the joint null hypothesis that the forward rate is an unbiased predictor of the future spot rate and that traders have rational expectations, both α and β should be zero. By definition, $\varepsilon_{t+q,q}$ has a mean of zero and is uncorrelated with all elements of the econometrician's information set at time t. When $q > 1$, models such as (7.2) do not have uncorrelated errors. To see this, note that $\varepsilon_{t+q,q}$ consists of unobservable information at time t as well as new information which is observable but arrives between time t and time $t + q$. Thus $\varepsilon_{t+q,q}$ is correlated with $\varepsilon_{t+q-a,q}$ for $1 \le a < q$, and $\{\varepsilon_{t+q,q}\}$ follows an MA($q - 1$). Further, as this new information arrives it becomes incorporated in the forward price so that $f_{t+a,q}$ is correlated with $\varepsilon_{t+q,q}$ for $1 \le a < q$. The random variable $f_{t,q}$ is therefore not exogenous but rather predetermined at time t.

Estimation of this equation can now proceed using OLS and the parameter estimators are consistent. They are not efficient however, since $\varepsilon_{t+q,q}$ follows an MA($q - 1$) process, and while this might appear to be a natural case for GLS, the nonexogeneity of our regressors presents a further problem. Let Ω represent the covariance matrix for $\varepsilon_{t+q,q}$ and let X be the $T \times 2$ design matrix corresponding to (7.2). Recalling that $f_{t+a,q}$ is correlated with $\varepsilon_{t+q,q}$, we see that $E[X'\Omega^{-1}\varepsilon] \ne 0$ and the GLS estimators are not consistent.

An important consideration in estimating models of this type is that the prediction errors, $s_{t+q} - f_{t,q}$, are not normally distributed. To see this we have constructed an omnibus test for nonnormality relying upon both skewness and kurtosis. We have chosen this test because we have no a priori knowledge of the suspected departures from normality exhibited by the data. In obtaining the critical values for the test statistic one must exercise some care owing to the dependence that arises between the two sample measures. This dependence can be accounted for by suitably adjusting the critical values of the test statistics as described in Pearson, D'Agostino, and Bowman (1977). For our sample size we have determined that the appropriate critical region for a nominal 5% test of the joint null hypothesis that $\sqrt{b_3}$ (the skewness measure) equals 0 and b_4 (the kurtosis measure) equals 3 is

constructed from the following two intervals:

$$(-0.698, 0.698) \text{ for } \sqrt{b_3}, \qquad (2.105, 4.450) \text{ for } b_4.$$

Our data consists of monthly spot rates and three-month-ahead forward rates collected by Barclay's Bank. Because of the broad cross-sectional composition of this data, it is available only from September of 1982 to January of 1988, which yields 65 sample points. This makes it appropriate for our study, as it is roughly comparable to the sample size explored in the Monte Carlo simulations of section 6.

For 16 of the 20 countries in our sample we can reject, at the 5% level, the null hypothesis that $s_{t+3} - f_{t,3}$ is normally distributed. The majority of these rejections indicate that the underlying distribution is platykurtic, with asymmetry appearing for three countries. It is interesting to note that in four cases, departures from normality were detected in the raw data but not in the dependent variable. For the majority of cases in which we can reject, this leads one to suspect that the $\varepsilon_{t+q,q}$ are nonnormal as well. In this context the adaptive estimators developed above are asymptotically fully efficient while OLS estimators are not. Furthermore, the efficiency gains occur in moderately sized samples so their application here seems warranted.

The results presented in table 3 compare the OLS estimates reported in MacArthur (1988) with estimates obtained using the adaptive estimators developed in section 4. The reported standard errors for the OLS estimates are calculated using a consistent estimate of the covariance matrix developed by White (1980). As noted in Steigerwald (1989b), the reported standard errors for the adaptive estimator are biased downward. This is a common feature in estimators with a nonparametric component, the effects of the smoothing parameter are not appropriately accounted for in the asymptotic variance expression.

To obtain more accurate standard errors we employ a bootstrap resampling device. Corresponding to our initial adaptive estimates of the parameters $(\hat{\alpha}, \hat{\beta}, \hat{\theta}_1, \hat{\theta}_2)$, we have the constructed white noise residuals $(\hat{u}_1, \ldots, \hat{u}_T)$. We then draw, with replacement, samples of 65 observations from the observed values of $(s_t - f_{t,q})$ and the estimated parameter values to form a constructed series for the dependent variable. Using this constructed data we can again estimate the parameters. This procedure is repeated 1000 times and the standard errors are calculated from the empirical distribution of the estimated parameters.

When looking at table 3 it should be pointed out that our data set is characterized by a strong appreciation of the dollar over much of the sampling period. Thus it may be the case that our rejections are due to a biased sample. For each of the open developed countries of the Atlantic,

Table 2

Omnibus tests for nonnormality.

		Skewness: $\sqrt{b_1}$	Kurtosis: b_2
Australia	Spot	0.15	2.16
	Forward	0.18	2.12
	$s_{t+3} - f_{t,3}$	1.12[a]	4.10
Belgium	Spot	−0.11	1.85[a]
	Forward	−0.12	1.86[a]
	$s_{t+3} - f_{t,3}$	−0.17	2.01[a]
Canada	Spot	−0.05	1.65[a]
	Forward	−0.02	1.69[a]
	$s_{t+3} - f_{t,3}$	−0.44	3.18
Denmark	Spot	−0.01	1.91[a]
	Forward	−0.02	1.94[a]
	$s_{t+3} - f_{t,3}$	−0.14	1.99[a]
France	Spot	0.23	2.02[a]
	Forward	0.17	2.02[a]
	$s_{t+3} - f_{t,3}$	−0.11	1.86[a]
Germany (W)	Spot	−0.10	1.89[a]
	Forward	−0.09	1.92[a]
	$s_{t+3} - f_{t,3}$	−0.11	1.96[a]
Greece	Spot	−0.81[a]	2.27
	Forward	−0.79[a]	2.28
	$s_{t+3} - f_{t,3}$	0.11	2.28
Hong Kong	Spot	−2.06[a]	5.92[a]
	Forward	−1.98[a]	5.96[a]
	$s_{t+3} - f_{t,3}$	1.74[a]	9.47[a]
Italy	Spot	0.36	2.03[a]
	Forward	0.31	2.04[a]
	$s_{t+3} - f_{t,3}$	−0.14	2.04[a]
Japan	Spot	−0.35	1.52[a]
	Forward	−0.36	1.54[a]
	$s_{t+3} - f_{t,3}$	−0.37	2.35
Malaysia	Spot	0.03	1.70[a]
	Forward	0.31	2.12
	$s_{t+3} - f_{t,3}$	0.14	2.58
Netherlands	Spot	0.15	1.94[a]
	Forward	0.17	1.98[a]
	$s_{t+3} - f_{t,3}$	0.12	1.97[a]
Norway	Spot	0.71[a]	2.18
	Forward	0.64	2.10
	$s_{t+3} - f_{t,3}$	0.55	2.07[a]
New Zealand	Spot	0.12	2.02[a]
	Forward	0.08	1.94[a]
	$s_{t+3} - f_{t,3}$	0.15	1.98[a]
Saudi Arabia	Spot	−0.89[a]	2.24
	Forward	−0.92[a]	2.18
	$s_{t+3} - f_{t,3}$	−0.84[a]	2.29

Table 2 (continued)

		Skewness: $\sqrt{b_1}$	Kurtosis: b_2
Singapore	Spot	0.24	1.86[a]
	Forward	0.32	1.99[a]
	$s_{t+3} - f_{t,3}$	0.20	1.93[a]
Spain	Spot	0.46	1.98[a]
	Forward	0.52	1.86[a]
	$s_{t+3} - f_{t,3}$	0.38	1.73[a]
South Africa	Spot	− 0.26	2.02[a]
	Forward	− 0.33	2.08[a]
	$s_{t+3} - f_{t,3}$	− 0.18	2.06[a]
Sweden	Spot	− 0.17	1.95[a]
	Forward	− 0.23	2.04[a]
	$s_{t+3} - f_{t,3}$	− 0.34	2.08[a]
Switzerland	Spot	0.36	1.64[a]
	Forward	0.54	1.53[a]
	$s_{t+3} - f_{t,3}$	0.47	2.19
United Kingdom	Spot	− 0.11	1.84[a]
	Forward	− 0.17	1.73[a]
	$s_{t+3} - f_{t,3}$	− 0.06	1.96[a]

[a]Significant at 5% level.

MacArthur found convincing evidence against the null hypothesis that α and β are both equal to 0. Of the remaining European developed countries (DC's) only Sweden provided a rejection. For the newly liberalizing developed countries of the Pacific, in two of the three cases the null hypothesis could again be rejected. However for the lesser developed countries (LDC's), including the liberalizing members of the Pacific basin, the null hypothesis could not be rejected for any of them. This led MacArthur to conclude that capital markets in these countries behaved in an empirically distinct way from their developed country counterparts. Further, this disparity could not be attributed solely to the fact that they were newly liberalizing countries since the liberalizing Pacific developed countries looked fairly similar to the open Atlantic developed countries.

Using the more precise adaptive estimators we confirm the rejections of the null hypothesis for the open Atlantic developed countries. For the remaining European developed countries, the adaptive estimators are more precise than their least squares counterpart in each case. For only two countries, France and Spain, are the gains in efficiency sufficient to reverse MacArthur's findings. However, for the rest of the sample our results are strikingly different and we find little evidence to distinguish lesser developed countries from the developed economies. In the Pacific region, we find that two-thirds of the LDC's reject the hypothesis of forward rate unbiasedness equivalent to the rejection frequency for the DC's. We even find evidence against this null hypothesis for one of the closed economy LDC's in our

Table 3

Estimated values for $s_{t+3} - f_{t,3} = \alpha + \beta(s_t - f_{t,3})$; standard errors in parentheses.

	OLS		Adaptive	
	α	β	α	β
Liberalizing Pacific DC's				
Australia	10.29	2.01	8.65	1.86
	(8.82)	(1.23)	(7.26)	(0.78)
Japan	− 38.02	9.58	− 37.42	8.63
	(11.17)	(3.31)	(11.84)	(3.64)
New Zealand	20.82	2.91	20.79	3.36
	(10.38)	(0.89)	(7.39)	(0.65)
Liberalizing Pacific LDC's				
Hong Kong	3.36	0.63	2.43	0.94
	(2.98)	(1.39)	(1.16)	(0.38)
Malaysia	1.57	1.03	1.48	1.01
	(2.06)	(0.36)	(2.08)	(0.43)
Singapore	3.68	− 1.59	3.73	− 1.84
	(3.61)	(1.64)	(1.73)	(0.88)
Closed LDC's				
Greece	9.13	0.85	9.17	0.77
	(9.02)	(0.57)	(10.01)	(0.56)
Saudi Arabia	1.57	0.79	1.39	0.98
	(0.50)	(0.45)	(0.48)	(0.25)
South Africa	3.97	− 0.13	6.53	− 0.28
	(9.62)	(0.29)	(5.07)	(0.11)
Open Atlantic DC's				
Canada	− 4.89	5.14	− 5.04	4.87
	(1.89)	(1.14)	(2.03)	(1.20)
Germany (W)	51.47	13.66	50.88	13.73
	(8.37)	(1.79)	(6.42)	(1.17)
Netherlands	38.04	12.17	37.60	9.63
	(7.29)	(1.63)	(4.91)	(1.26)
Switzerland	44.89	9.39	42.87	9.47
	(11.43)	(2.27)	(12.66)	(2.63)
United Kingdom	− 12.99	9.25	− 13.06	11.12
	(3.07)	(1.35)	(2.58)	(1.06)
Other European DC's				
Belgium	6.53	1.99	5.89	2.84
	(6.58)	(3.99)	(4.83)	(1.60)
Denmark	6.33	1.00	6.75	1.72
	(6.61)	(1.91)	(3.24)	(0.94)
France	9.82	− 0.73	9.88	− 1.36
	(7.45)	(1.17)	(4.62)	(0.63)

Table 3 (continued)

	OLS		Adaptive	
	α	β	α	β
Italy	5.71	0.51	7.83	0.48
	(10.28)	(1.23)	(7.44)	(0.96)
Norway	− 13.57	4.27	− 13.98	1.80
	(9.15)	(1.49)	(7.61)	(1.13)
Spain	13.35	− 0.53	11.61	− 0.77
	(8.81)	(0.76)	(4.66)	(0.28)
Sweden	− 12.00	6.04	− 12.36	4.21
	(4.66)	(1.03)	(4.21)	(0.87)

sample, Saudi Arabia. In summary, using our potentially more efficient estimators, we find little evidence that capital markets behaved differently in the developed countries and the lesser developed countries over the 1980's.

There may be a simple explanation for these results. One of the features common to both the less developed countries and the newly liberalizing developed countries is a recently created international financial marketplace. Because of their relatively short history, we might expect data from these markets to be characterized by a high degree of 'noise' resulting from the still inefficient market channels through which information flows. In using an estimation technique that is capable of screening out such excess noise, one should obtain better estimates of the relationship between spot rates and forward rates, and therefore results that are more in line with those of the economies with full integrated international financial markets. The adaptive estimators provide results for countries with newly developing international capital markets that accord with those obtained for the traditional open economies.

8. Conclusion

The above results demonstrte that it is possible to construct estimators for the parameters of a regression model that are asymptotically fully efficient without knowledge of the underlying error sequence. These results pertain to models in which the error process is serially correlated and the asymptotic results are unaffected when the order of the ARMA process is unknown. The Monte Carlo study indicates that substantial efficiency gains are possible in small samples and our investigation of forward premia on foreign currency markets supports this finding.

In addition to the practical applications highlighted in this paper, several theoretical questions indicate useful extensions of this work. One concerns the applicability of this technique to situations in which the errors do not

have a symmetric distribution. Initial work indicates that it may be possible to determine even weaker conditions under which adaptive estimators can be constructed. This work also appears to have natural extensions to multivariate problems. In constructing the likelihood function, the estimated score would involve a nonparametric estimate of the joint density of the errors from the equations in the model.

Appendix

The expression that we use for $u_t^T - u_t$ is an extension of the notation originally developed in Kreiss (1987). Under Assumption 2, the moving average polynomial can be inverted and the coefficients from the power series expansion satisfy the recursion

$$A_i + A_{i-1}\theta_1 + \cdots + A_{i-q}\theta_q = 0,$$

where $A_0 = 1$ and $A_i = 0$ if $i < 0$. Using this recursion and setting $\rho_0 = \theta_0 = 1$, we have

$$u_t = \sum_{i=0}^{t-1} A_i \sum_{j=0}^{p} \rho_j \varepsilon_{t-i-j} + \sum_{s=0}^{q-1} u_{-s} \sum_{j=0}^{s} A_{t+s-j}\theta_j.$$

From this, simple algebra yields (3.2) with

$$\mathscr{S}_t(\alpha^T, \alpha) = \sum_{i=0}^{t-1} A^T[\mathscr{S}_{1,t}' : \mathscr{S}_{2,t}' : \mathscr{S}_{3,t}'], \tag{A.1}$$

where

$$\mathscr{S}_{1,t} = -T^{-1/2} \sum_{j=0}^{p} \rho_j[x_{1,t-i-j}, \ldots, x_{k,t-i-j}],$$

$$\mathscr{S}_{2,t} = T^{-1/2}[\varepsilon_{t-i}, \ldots, \varepsilon_{t-i-p}],$$

$$\mathscr{S}_{3,t} = T^{-1/2}[u_{t-i}, \ldots, u_{t-i-q}].$$

Proof of Theorem 4.1

Using the definition of $\hat{\alpha}^T$ given in (4.5), the left-hand side of (4.1) can be rewritten as

$$T^{1/2}(\bar{\alpha}^T - \alpha) + \mathscr{I}(\alpha)^{-1}\left[\mathscr{I}(\alpha)\hat{\mathscr{I}}_T(\bar{\alpha}^T)^{-1}\hat{S}_T(\bar{\alpha}^T) - S_T(\alpha)\right],$$

$$\tag{A.2}$$

where $\hat{S}_T(\bar{\alpha}^T)$ and $\hat{\mathscr{I}}_T(\bar{\alpha}^T)$ incorporate estimates of f and are defined in (4.6) and (4.7).

In proving that this converges in probability to 0, we begin by noting the ξ_t are independent and identically distributed random variables, P_{α^T} and P_α are contiguous, and $\bar{\alpha}^T$ is $T^{1/2}$-consistent, so, using Chebychev's WLLN,

$$T^{-1} \sum_{t=1}^{T} \xi^2\big[u_t(\bar{\alpha}^T)\big] \to I(f),$$

in probability under α. Therefore we need only prove that

$$T^{-2} \sum_{t=1}^{T} \mathrm{E}\big\|\mathscr{I}_t(\bar{\alpha}^T)\big\|^2 \int \big[q(z,\hat{f}) - \xi(z)\big]^2 f(z)\,\mathrm{d}z \qquad (\mathrm{A}.3)$$

converges to 0. In proving that our estimate of ξ converges in quadratic mean to its expected value, we will follow the logic employed by Bickel (1982). We can bound $\int[q(z,\hat{f}) - \xi(z)]^2 f(z)\,\mathrm{d}z$ by

$$3\bigg(\int \Big\{ q\big(z, \hat{f}_{\sigma,t}\big) - q\big(z, \hat{f}_{\sigma,t}\big)\big[\sqrt{f_\sigma(z)} \big/ \sqrt{f(z)} \big] \Big\}^2 f(z)\,\mathrm{d}z$$

$$+ \int \Big\{ q\big(z, \hat{f}_{\sigma,t}\big)\big[\sqrt{f_\sigma(z)} \big/ \sqrt{f(z)} \big]$$

$$- \big[\dot{f}_\sigma(z)/f_\sigma(z) \big]\big[\sqrt{f_\sigma(z)} \big/ \sqrt{f(z)} \big] \Big\}^2 f(z)\,\mathrm{d}z$$

$$+ \int \Big\{ \big[\dot{f}_\sigma(z)/f_\sigma(z) \big]\big[\sqrt{f_\sigma(z)} \big/ \sqrt{f(z)} \big] - \xi(z) \Big\}^2 f(z)\,\mathrm{d}z \bigg), \quad (\mathrm{A}.4)$$

where $f_\sigma(z)$ represents the convolution of $f(z)$ and the density of ξ corresponding to the value of σ used in constructing $\hat{f}_{\sigma,t}$. Combining (A.3) and (A.4) we have three terms. For the first and third we can rely upon the result, given in Steigerwald (1989a), that $\mathrm{E}\|\mathscr{I}_t(\bar{\alpha}^T)\|^2$ is uniformly bounded in T. Therefore, when $\sigma_T \to 0$ and $g_{3,T}\sigma_T \to 0$, we can use Lemmas 6.2 and 6.3 from Bickel to show that these terms converge to zero.

For the second term let C, D, and H represent the conditions given in (4.4), then, using the relationships noted in Stone (1975), we can place the

following bound on $\int [q(z, \hat{f}_{\sigma,t}) - \dot{f}_\sigma(z)/f_\sigma(z)]^2 f_\sigma(z) \, dz$:

$$2 \left\{ \int_{\text{CDH}} g_{3,T}^2 \operatorname{var}\left[\dot{f}_{\sigma,t}(z)\right] f_\sigma^{-1}(z) \, dz \right.$$

$$\left. + \int_{\text{CDH}} g_{3,T}^2 \operatorname{var}\left[\hat{\dot{f}}_{\sigma,t}(z)\right] f_\sigma^{-1}(z) \, dz \right\}. \tag{A.5}$$

Using (A.5) to construct the second term, we note that X is independent of $\hat{f}_{\sigma,t}(z)$ by assumption. Combining this with the fact that f has a finite fourth moment implies that the second term converges in probability to 0 whenever $g_{1,T}\sigma_T^{-3}$ is o(T).

Proof of Corollary 5.2

We must alter our notation slightly to account for the fact that we are estimating the parameters p and q. Let $\hat{p}^T$ and $\hat{q}^T$ represent the estimators constructed from the Lagrange multiplier (LM) test procedure based upon a sample size T. Our estimator of the score incorporates $\mathscr{S}_t(\bar{\alpha}^T, \hat{p}^T, \hat{q}^T)$, where

$$\mathscr{S}_t(\bar{\alpha}^T, \hat{p}^T, \hat{q}^T)$$

$$= \sum_{h=0}^{t-1} A_h^T \left[-X'_{t-h} + \cdots + \rho_{\hat{p}^T}^T X'_{t-h-\hat{p}^T} \mid \varepsilon_{t-1-h}, \ldots, \varepsilon_{t-\hat{p}^T-h} \mid \right.$$

$$\left. u_{t-1-h}, \ldots, u_{t-\hat{q}^T-h} \right].$$

In Steigerwald (1989a) we have shown that when f is known, $G_T(\bar{\alpha}^T) \to G(\alpha)$ and $\mathscr{I}(\alpha)T^{1/2}(\bar{\alpha}^T - \alpha) + S_T(\bar{\alpha}^T) - S_T(\alpha) \to 0$ guaranteeing (4.1). Consistency of our estimators $\hat{p}^T$ and $\hat{q}^T$ implies

$$P(\hat{p}^T = p, \hat{q}^T = q) \to 1,$$

as $T \to \infty$. Employing the first lemma in Pötscher (1990), we have

$$P(G_T(\bar{\alpha}^T, \hat{p}^T, \hat{q}^T) = G_T(\bar{\alpha}^T)) \to 1,$$

$$P(S_T(\alpha, \hat{p}^T, \hat{q}^T) = S(\alpha)) \to 1,$$

both as $T \to \infty$. Thus our asymptotic results hold and the estimator $\hat{\alpha}_T$ remains LAM when p and q are unknown.

Proof of Theorem 5.3

First observe that our LM test statistic uses as its criterion function the squared prediction error. It is independent of the underlying density and does not require knowledge of f for its construction. Thus consistent estimation remains possible when f is unknown.

We only need show that our estimator $q(z, \hat{f}_{\sigma,t}, \hat{p}^T, \hat{q}^T)$ is asymptotically equivalent to $q(z, \hat{f}_{\sigma,t}, p, q)$. Using the logic employed above we can see that

$$P\left(q\left(z, \hat{f}_{\sigma,t}, \hat{p}^T, \hat{q}^T\right) = q\left(z, \hat{f}_{\sigma,t}, p, q\right)\right) = 1,$$

as $T \to \infty$ for all elements t.

References

Beran, R. 1976, Adaptive estimates for autoregressive processes, Annals of the Institute of Statistical Mathematics 28, 77–89.

Bickel, P., 1982, On adaptive estimation, Annals of Statistics 10, 647–671.

Bollerslev, T., R. Chou, and K. Kroner, 1992, ARCH modeling in finance: A review of the theory and empirical evidence, Journal of Econometrics 52, 5–59.

Chamberlain, G., 1986, Asymptotic efficiency in semiparametric models with censoring, Journal of Econometrics 32, 189–218.

Coslett, S., 1983, Distribution-free maximum likelihood estimator of the binary choice model, Econometrica 51, 765–782.

Cox, J., J. Ingersoll, and S. Ross, 1981, The relation between forward prices and futures prices, Journal of Financial Economics 9, 321–346.

Deistler, M., 1983, The properties of the parameterizations of ARMAX systems and their relevance for structural estimation and dynamic specification, Econometrica 51, 1187–1207.

Duncan, G., 1986, A semiparametric censored regression estimator, Journal of Econometrics 32, 5–34.

Fabian, V. and J. Hannan, 1982, On estimation and adaptive estimation for locally asymptotically normal families, Zeitschrift für Wahrscheinlichkeitstheorie 59, 459–478.

Grendander, U., 1954, On the estimation of regression coefficients in the case of an autocorrelated disturbance, Annals of Mathematical Statistics 25, 252–272.

Hajek, J., 1972, Local asymptotic minimax and admissibility in estimation, in: L. Le Cam, J. Neyman, and E. Scott, eds., Proceedings of the sixth Berkeley symposium on mathematical statistics and probability (University of California Press, Berkeley, CA) 175–194.

Hsieh, D. and C. Manski, 1987, Monte Carlo evidence on adaptive maximum likelihood estimation of a regression, Annals of Statistics 15, 541–551.

Kreiss, J., 1987, On adaptive estimation in stationary ARMA processes, Annals of Statistics 15, 112–133.

Le Cam, L., 1986, Asymptotic methods in statistical decision theory (Springer-Verlag, New York, NY).

MacArthur, A., 1988, Tests of forward discount unbiasedness with 25 currencies, Doctoral dissertation (University of California, Berkeley, CA).

Manski, C., 1984, Adaptive estimation of nonlinear regression models, Econometric Reviews 3, 145–194.

Manski, C., 1987, Semiparametric analysis of random effects linear models from binary panel data, Econometrica 55, 357–362.

Newey, W., 1986, Linear instrumental variable estimation of limited dependent variable models with endogenous explanatory variables, Journal of Econometrics 32, 127–141.

Newey, W., 1990, Semiparametric efficiency bounds, Journal of Applied Econometrics 5, 99–135.

Pearson, E., R. D'Agostino, and K. Bowman, 1977, Tests for departure for normality: Comparison of powers, Biometrika 64, 231–246.

Pötscher, B., 1983, Order estimation in ARMA models by Lagrangian multiplier tests, Annals of Statistics 11, 872–885.

Pötscher, B., 1990, Effects of model selection on inference, Manuscript (University of Maryland, College Park, MD).

Powell, J., 1986, Symmetrically trimmed least squares estimation for Tobit models, Econometrica 54, 1435–1460.

Ruud, P., 1986, Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models, Econometrica 51, 225–228.

Steigerwald, D., 1989a, Adaptive estimation in time series models, Manuscript (University of California, Santa Barbara, CA).

Steigerwald, D., 1989b, Monte Carlo evidence on the efficiency gains of adaptive estimators, Journal of Econometrics, forthcoming.

Stein, C., 1956, Efficient nonparametric testing and estimation, in: J. Neyman, ed., Proceedings of the third Berkeley symposium on mathematical statistics and probability (University of California Press, Berkeley, CA) 187–195.

Stein, C., 1960, Multiple regression, in: I. Olkin et al., eds., Contributions to probability and statistics (Stanford University Press, Stanford, CA) 424–443.

Stone, C., 1975, Adaptive maximum likelihood estimators of a location parameter, Annals of Statistics 3, 267–284.

Swensen, A., 1985, The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend, Journal of Multivariate Analysis 16, 54–70.

White, H., 1980, A heteroscedasticity consistent covariance matrix estimator and a direct test for heteroscedasticity, Econometrica 48, 817–838.