Journal of Econometrics 54 (1992) 371–400. North-Holland

On the finite sample behavior of adaptive estimators

Douglas G. Steigerwald*
University of California, Santa Barbara, CA 93106, USA

Received December 1989, final version received October 1991

We explore the finite sample performance of adaptive estimators in linear time series regression models. Our results show that for samples of only 50 observations, the 90 percent confidence interval of the adaptive estimator is 20 to 50 percent smaller than the corresponding interval for its GLS counterpart across a range of symmetrical distributions. When the assumption of symmetry is relaxed smaller gains are observed. These results are sensitive to the degree of departure from normality and the precision of the measurement exercise. We further observe that the estimated standard errors are biased downward.

1. Introduction

Econometric modelling has been increasingly influenced by developments in nonparametric statistics. This is reflected in the growing popularity of distribution-free estimators. However, most of the research in this area has been directed at the asymptotic properties of these estimators with little attention given to their finite sample properties. This paper attempts to redress the imbalance by comparing the small sample behavior of several estimators under a variety of distributional assumptions.

We focus upon a type of semiparametric regression model in which the regression function is specified exactly while the distribution of the errors is assumed only to lie within a broad class of distribution functions. To understand the generality of this approach consider the linear regression model in which there is a fixed set of independent variables and the error

*I would like to thank Takeshi Amemiya, Kjell Doksum, Ed Leamer, Dick Meese, Tom Rothenberg, Paul Ruud, participants at numerous seminars and the Econometric Society 1991 Summer Meetings, and two anonymous referees for helpful comments. I would also like to thank the Cray Corporation for generous financial assistance. All computations were done on the Cray X-MP at U.C. Berkeley and at the University of Illinois.

density is centered at the origin. Least squares estimators of the slope coefficients will be consistent and asymptotically normal under a wide range of distributions. They will not, in general, be asymptotically efficient. Other semiparametric estimators exist which are asymptotically efficient under all densities with a finite fourth moment. Such densities encompass a wide range of distributions including both the familiar Student's t-distribution, a large family of mixtures of normal random variables, and lognormal distributions.

The asymptotic efficiency bound which a semiparametric estimator attains, often called the semiparametric efficiency bound, can be compared with the asymptotic efficiency bound associated with the maximum likelihood estimator. The difference between these two is a measure of the efficiency loss associated with the more general specification of the error distribution admitted by the semiparametric estimator.

Our attention will be focused on situations in which the efficiency bound of the semiparametric estimator equals that of the maximum likelihood estimator. In these cases the intercept and slope coefficients can be estimated adaptively. The term adaptive embodies the notion that these estimators adapt to the sample by using the data to nonparametrically estimate a function of the density. Adaptive estimators can be constructed for many of the models used in economics and, as with all estimators, an understanding of their finite sample behavior is necessary before they can be applied with confidence. Knowledge of small sample properties is especially important in adaptive estimation, since the nonparametric estimator of the density requires several smoothing and trimming parameters over which the theory provides no finite sample guidance.

In examining the finite sample behavior of adaptive estimators Hsieh and Manski (1987) present some promising results for the linear model with serially uncorrelated errors. Their Monte Carlo study, conducted across a range of distributions with a sample size of 50, yields an interquartile range for the adaptive estimator of the slope coefficient that is 30 percent smaller than the corresponding range for the ordinary least squares estimator. Further, when the sample size was reduced to only 25 observations the adaptive estimator continued to outperform its ordinary least squares counterpart. They also determined that while the optimal choices of the trimming parameters were not sensitive to the specified distribution, the choice of the smoothing parameter unfortunately was. Since the attractiveness of adaptive estimators lies in the weak distributional assumptions used to derive them, this could present a barrier to their application in empirical work. To overcome this problem we allow the data to determine directly the choice of the smoothing parameter. This is accomplished by choosing the smoothing parameter to minimize a risk function corresponding to the mean integrated squared error. While the methodology differs from that of Hsieh and Manski, the results are not quantitatively different.

In our study we consider a more comprehensive family of models so that a researcher may use it as a guide to determine when efficiency gains are possible. We use a linear regression framework but introduce serial correlation in the errors. Since serially correlated errors are often associated with serially correlated regressors, we also study a model in which the independent variables are autocorrelated.

The paper is organized as follows. The next section presents a brief review of adaptive estimators. Section 3 describes the Monte Carlo experiments which are performed. The fourth section describes the results of those experiments, and section 5 presents concluding remarks.

2. Adaptive estimators

The theory defining adaptive estimators was first developed by Stein (1956) and then extended by Bickel (1982). While Bickel's work incorporated models of interest to econometricians it did not allow for serial dependence of the errors. Adaptive estimators in linear regression models with serially correlated errors are treated formally in Steigerwald (1990, 1992) where, unlike the work of Bickel, sample splitting is not required. These results can be easily extended to nonlinear regression models following the work of Manski (1984). A brief review of this literature should provide the reader with an appropriate frame of reference.

Consider the linear regression model

$$y_t = x_t'\beta + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{1}$$

in which the errors are characterized by an autoregressive moving average process (ARMA) of order (p, q)

$$\varepsilon_t = u_t + \sum_{i=1}^{q} \theta_i u_{t-i} + \sum_{j=1}^{p} \rho_j \varepsilon_{t-j}. \tag{2}$$

The random variables, $\{u_t\}$, are assumed to be independent, the ARMA process is assumed to be stationary and invertible, and its order is known. When the order of the ARMA process is unknown, a consistent estimator of the order is all that is needed to obtain adaptive estimators [see Steigerwald (1992)].

If we are interested in estimating $\omega = [\beta', \rho', \theta']$, then we may treat the other parameters characterizing the density of u as a vector of nuisance parameters. To understand how one can estimate ω efficiently without specifying the distribution, let us begin with the case in which the density of u, $f^*(\cdot)$, is known up to the vector η. The likelihood function, $L(y|x, \omega, \eta)$,

can then be used to form Fisher's information matrix, the inverse of which provides our asymptotic efficiency bound. When the information matrix is block-diagonal,

$$\mathrm{E}[\partial \ln L(y|x,\omega,\eta)/\partial\omega\, \partial \ln L(y|x,\omega,\eta)/\partial\eta'] = 0, \tag{3}$$

and the asymptotic efficiency bound for ω is not affected by the presence of η.

If we now assume that $f^*(\cdot)$ is known only to lie in some convex space of functions, F, then the asymptotic variance of our estimator of ω must be at least as large as when $f^*(\cdot)$ was assumed to lie in a subfamily of F characterized by the parameter η. Moreover, an adaptive estimator of ω must satisfy an orthogonality condition that is analogous to (3). We can now think of the nuisance vector as being infinite-dimensional. The score function for ω, the partial derivative of the log-likelihood function with respect to ω, will be orthogonal to the score for this nuisance vector if the following two conditions hold.

First, given the true score function, its expectation must be zero over the entire class F. Thus for any f an element of F,

$$\mathrm{E}_f[\partial \ln L(y|x,\omega,f^*)/\partial\omega] = 0. \tag{4}$$

Second, it must be possible to construct a nonparametric estimator of the score function which converges to the true score function in quadratic mean. Steigerwald (1990) shows that these conditions are met for the class of densities centered at the origin with finite fourth moments.

The resultant estimators are two-step estimators using an initial $\sqrt{T}$-consistent estimate as the starting value. Since the method of least squares provides consistent estimators in the models we will study, it is used to form the initial estimate. The adaptive estimator is then analogous to the linearized likelihood estimator (LLE), where the score function and information matrix are replaced with their nonparametric estimates. We know that one-step LLE estimators are asymptotically fully efficient and that one-step estimators remain consistent if we replace the true score with a fixed estimate of it. Adaptive estimators use an estimate of the score that converges to the true score, regaining full asymptotic efficiency.

The nonparametric estimate of the density is constructed using the residuals from the initial least squares regression. The density estimate is calculated for each of these values using a cross-validated normal kernel. For each value of the residuals, a distance measure is calculated between that residual and each of the other residuals. A normal probability density function is then evaluated at each of these distance measures and is averaged to produce the estimate of the density. Since we will be examining a derivative of this density

estimate, we need to define it over a small neighborhood around each of our residuals, e_t. Let our estimator of the density at e_t be defined for all z in a small neighborhood of each e_t as

$$\hat{f}_t(z) = [(T-1)]^{-1} \sum_{s=1,\, s\neq t}^{T} \xi_\sigma(z - e_s), \tag{5}$$

where ξ_σ is the probability density function corresponding to a normal random variable with mean zero and variance σ^2. This variance controls the amount of smoothing, as σ^2 decreases the weights given to residuals which lie some distance from e_t tend to zero.

Let the nonparametric estimator of the derivative of the density, $\hat{\hat{f}}_t(\cdot)$, be the derivative of the function given in (5), and define the nonparametric estimator of the score function as

$$\begin{aligned}\hat{S}(\omega) &= \partial \ln L\big(y|x, \omega, \hat{f}\big)/\partial\omega \\ &= T^{-1} \sum_{t=1}^{T} \partial e_t/\partial\omega \Big[\hat{\hat{f}}(e_t)/\hat{f}(e_t)\Big].\end{aligned} \tag{6}$$

We restrict the behavior of this estimator using the following three trimming conditions. We set the score equal to zero if either the value of the residual is too large,

$$|e_t| \geq tr_1, \tag{7a}$$

the value of the density estimate is too small,

$$\hat{f}(e_t) < tr_2, \tag{7b}$$

or the value of the 'updating step' is too large,

$$\big|\hat{\hat{f}}(e_t)\big|\big/\hat{f}(e_t) > tr_3. \tag{7c}$$

We then use the outer product of the score function to estimate the information matrix. If we let $\bar{\omega}$ denote the OLS estimator, then our adaptive estimator, $\hat{\omega}$, is given by

$$\hat{\omega} = \bar{\omega} + \left[T^{-1} \sum_{t=1}^{T} \partial \ln L\big(y|x, \bar{\omega}, \hat{f}\big)/\partial\omega\, \partial \ln L(\cdot)/\partial\omega'\right]^{-1} \hat{S}(\bar{\omega}). \tag{8}$$

3. Model design

The Monte Carlo experiments are designed to provide researchers with a guide to the potential gains arising from the application of adaptive estimators. They are based upon a linear model in which the error term follows several alternative ARMA processes and the white noise residuals are generated under a variety of distributional assumptions. In addition, since many models in economics are characterized by serial correlation in both the regressors and the errors, we consider both serially correlated and independent sequences of exogenous variables.

Our basic linear model is given by

$$y_t = \alpha + \beta x_t + \varepsilon_t, \tag{9}$$

where $\alpha = -1$, $\beta = 1$, and x_t is statistically independent of ε_t. While it would be of interest to explore more complicated models to monitor the performance of the estimators, the resulting family of models is extremely broad, and any selected member is subject to the criticism that it is not representative of models typically encountered in practice. A sequence of experiments for a group of more complicated models would be extremely costly to perform; in fact, merely increasing the dimension of the independent variables in the above model is prohibitively expensive. We therefore restrict attention to the simple linear regression model given in (9) and compensate for this by including an extremely small sample size, $T = 50$.

In the first class of experiments the error is assumed to be independent through time and ε_t is set equal to the white noise random variable, u_t. The second group of experiments introduces serial correlation in the errors through a stationary, invertible first-order autoregression [AR(1)],

$$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t. \tag{10}$$

The third class of experiments replaces the AR(1) with an invertible first-order moving average process [MA(1)],

$$\varepsilon_t = u_t + 0.5u_{t-1}. \tag{11}$$

While this remains a one-parameter error process, the moving average parameter is much more difficult to estimate accurately than the autoregressive parameter, providing additional information about the importance of the initial estimates of the error process.

For each of the above experiments, we allow the distribution of the white noise residuals to vary. We first examine the case in which u follows a normal distribution, to calculate the efficiency losses resulting from an unnecessary

Table 1

Characteristics of the density functions.[a]

Name	Construction	Mean	Var.	Skew.	Kurt.
Unimodal 1	$0.9N(0,1) + 0.1N(0,3)$	0	1.2	0	3.75
Unimodal 2	$0.9N(0,1) + 0.1N(0,10)$	0	1.9	0	9.06
Unimodal 3	$0.9N(0,1/9) + 0.1N(0,9)$	0	1.0	0	24.33
Bimodal 1	$0.5N(-1,1) + 0.5N(1,1)$	0	2.0	0	2.50
Bimodal 2	$0.5N(-3,1) + 0.5N(3,1)$	0	10.0	0	1.38
Bimodal 3	$0.5N(-10,1) + 0.5N(10,1)$	0	101.0	0	1.04
Lognormal 1	$\exp(z)$ where $z \sim N(0,0.01)$	1.01	0.01	0.30	3.16
Lognormal 2	$\exp(z)$ where $z \sim N(0,0.10)$	1.05	0.12	1.01	4.86
Lognormal 3	$\exp(z)$ where $z \sim N(0,1.0)$	1.65	4.67	6.18	113.94

Sample results of the tests of nonnormality

Distribution	Skew.	Kurt.
Unimodal 1		20%
Unimodal 2		68%
Unimodal 3		97%
Bimodal 1		22%
Bimodal 2		99%
Bimodal 3		99.9%
Lognormal 1	20%	13%
Lognormal 2	71%	35%
Lognormal 3	99.9%	94%

[a]All of the densities are standardized to have a mean of zero and a variance of 1 for the actual data generation.

weakening of assumptions. We then alter the shape of the distribution in a variety of ways, standardizing each distribution to have a mean of 0 and variance of 1.

The two most frequently described departures from normality are thick tails and asymmetry. We attempt to capture the empirical concept of thick tails using two classes of sequences of density shape deformations. The first is related to the statistical definition of contamination. We study three unimodal symmetric normal mixtures that are analogous to situations in which 90 percent of the time the data are generated by a given distribution, but 10 percent of the time they are drawn from a distribution with a larger variance. This corresponds, for example, to intermittent periods of noise in financial markets during which increased uncertainty on the part of investors leads to an increase in the measured variance of financial variables.

These three mixtures, and their corresponding moment characteristics, are listed in table 1 as unimodal 1 through 3. All three are mean 0 by construction, and the probability of obtaining an outlying value increases with the difference between the variances of the normal random variables. While each

of these distributions has more weight in the tails than a standard normal distribution, a cursory examination of fig. 1 reveals the increasing 'peakedness' of these mixtures. This is reflected by the increasing kurtosis summarized in table 1. A normal random variable has a kurtosis of 3. Distributions with a kurtosis that exceed 3 are leptokurtic and have a higher concentration of probability mass near their mode. Thus our first group of normal mixtures differs from a normal random variable in two important ways, they have thicker tails and more of their density mass is concentrated near the mean.

The second sequence of normal mixtures is also constructed under the hypothesis that the data is drawn from two different populations. The two populations have different means and are selected with equal probability, creating bimodal distributions with tails that are thicker than a normal distribution's. The moment properties of these bimodal mixtures are listed in the first table as bimodal 1 through 3. As the distance between the means grows the kurtosis falls toward 1, and this is reflected in fig. 2 by the decreasing height of the density function near 0. We see that this group of symmetric normal mixtures is also characterized by thicker tails than a normal distribution, yet these have less of their density mass concentrated near the mean.

Many empirical problems are characterized by skewed distributions, and we introduce this departure from normality in the form of the lognormal family. Our three skewed distributions, listed in table 1 as lognormal 1 through 3, exhibit an increase in both skewness and kurtosis as the variance of the underlying normal random variable increases. This can be readily seen in fig. 3, where the increasing asymmetry is associated with a more peaked density.

In applied problems researchers will not know the true underlying error distribution and must rely upon summary measures to detect departures from normality. It is of interest to see how sensitive these tests are to the densities we are using. Two of the most common tests measure the sample skewness and the sample kurtosis. The hypothesis of normality is rejected when either the first differs significantly from zero or the second differs significantly from three.

The top half of table 1 presents the relevant population values for our densities, and the bottom half presents the test results based upon the generated samples. For each sample, constructed under the assumption that the errors are independent, estimates of the skewness and kurtosis are obtained. The lower portion of the first table reports how frequently a test of the null hypothesis of a normal data-generating process is rejected at the 5 percent level. [The critical values of these tests for a sample of size 50 are given in White and MacDonald (1980).] For each of the symmetric distributions the rejection proportions are given for the kurtosis test, while for the

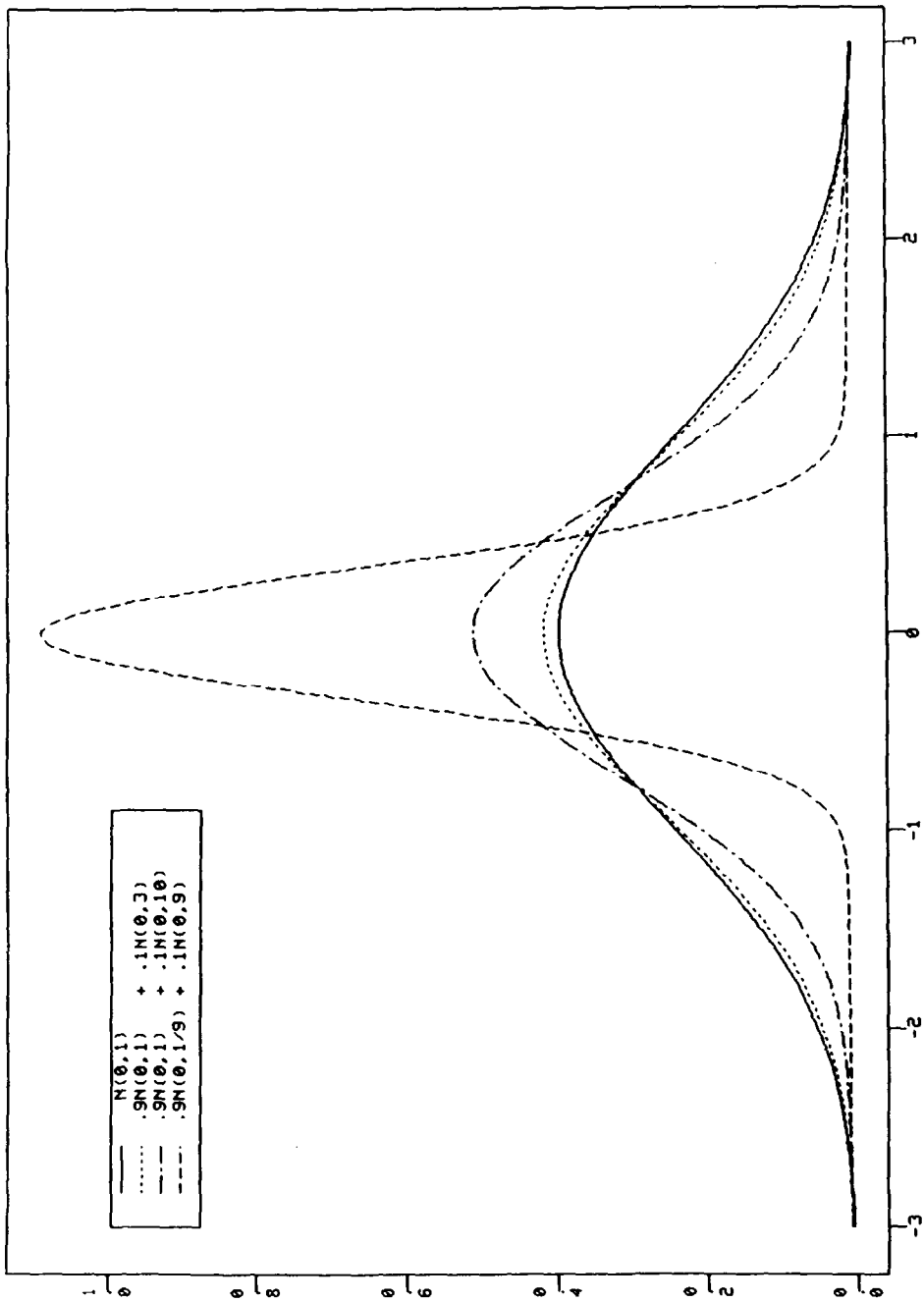

Fig. 1. Standardized unimodal symmetric normal mixtures.

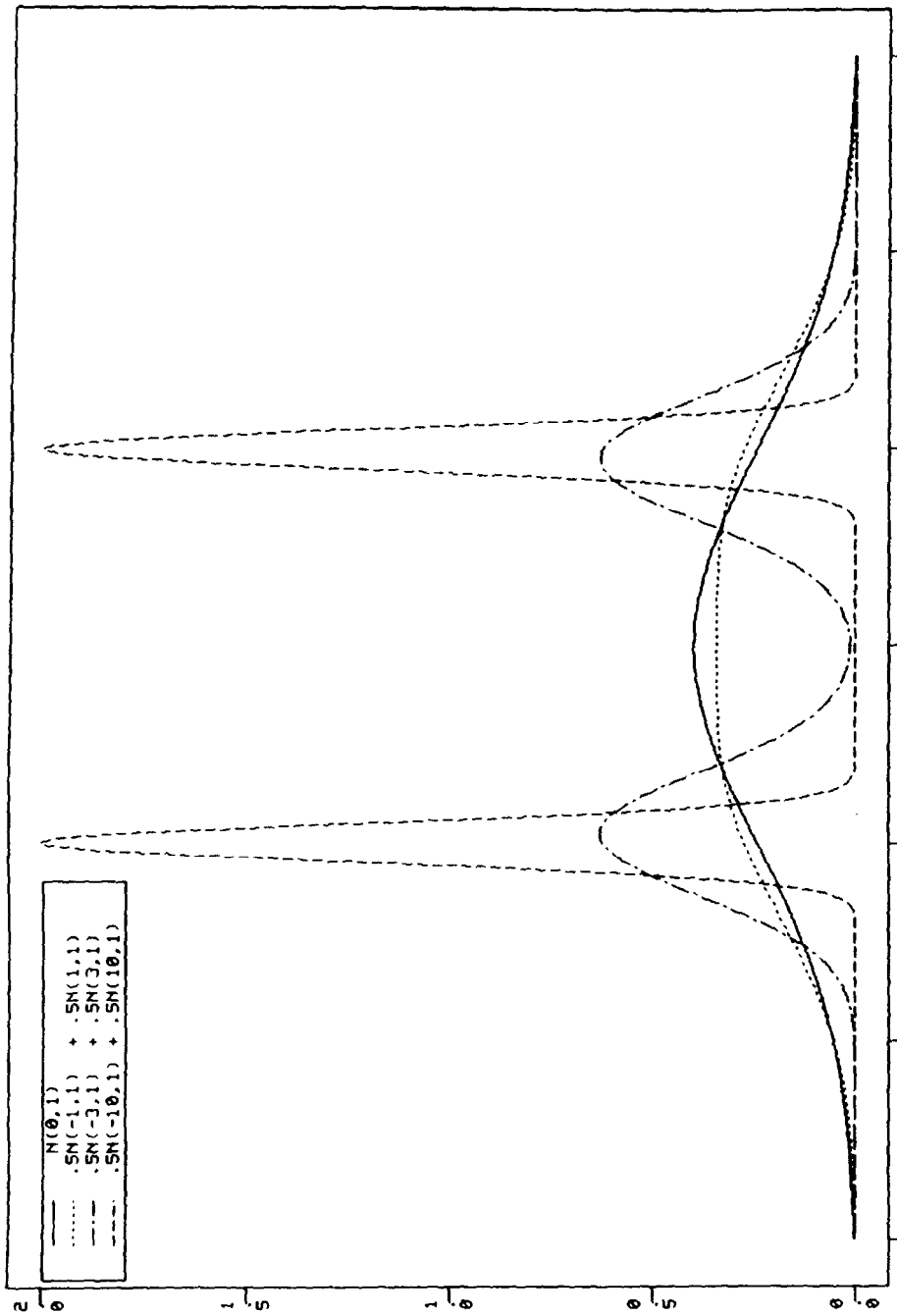

Fig. 2. Standardized bimodal symmetric normal mixtures.

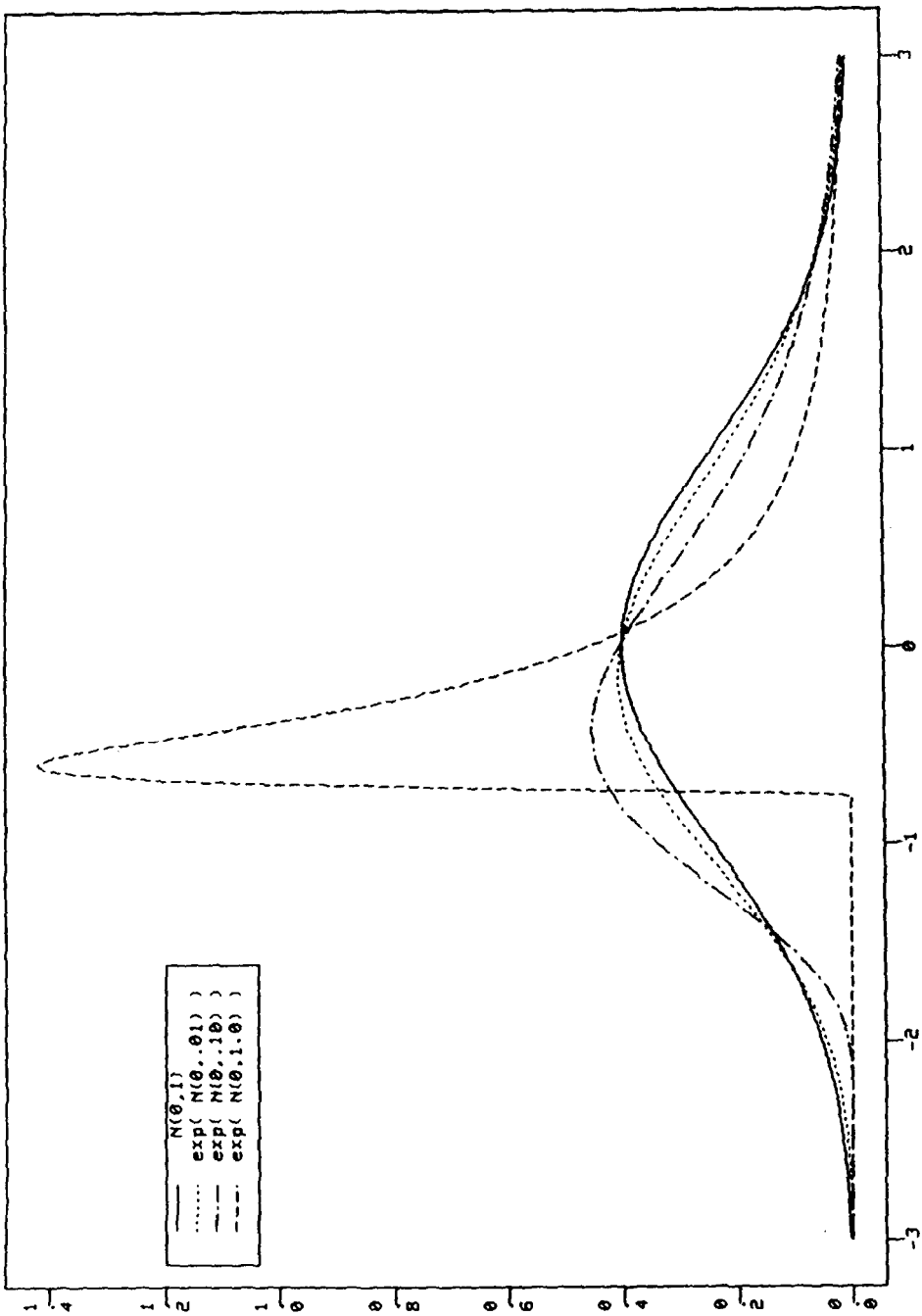

Fig. 3. Standardized lognormal densities.

lognormal distributions these rejection proportions are calculated for both the skewness and kurtosis tests.

For each combination of error process and white noise density, the exogenous variable is generated in two different ways. In one case x_t is a Bernoulli random variable with an equal probability of assuming either 0 or 1. In the second, x_t follows a stationary and invertible AR(1),

$$x_t = 0.7x_{t-1} + \eta_t, \tag{12}$$

where η_t is drawn from the standardized lognormal 3 distribution. While the second model for x_t incorporates serial correlation, a common property of the regressors when the error is serially dependent, it also increases the variation in x_t giving rise to the more precise estimation results in the experiments that follow.

A Monte Carlo experiment involves the selection of the error process, the white noise density, and the exogenous variable process. The initial estimate used to construct both the LLE and adaptive estimators is the OLS estimate. The initial consistent estimate of the autoregressive parameter is obtained by regressing ε_t on ε_{t-1}. For the moving average coefficient, the autocorrelation function is obtained and an iterative solution involving the autocovariance is used.

The values of the trimming parameters are chosen in such a way that when the underlying density is normal, all three are binding at the same value. This implies that $tr_1 = m$, $tr_2 = \exp(-m^2/2)$, and $tr_3 = m$. We have followed Hsieh and Manski in selecting $m = 8$ unless otherwise specified. While we fix the trimming parameters, we allow the data to select the smoothing parameter. Rather than adopt the bootstrap algorithm of Hsieh and Manski (1987) we choose to minimize the mean integrated squared error. When a normal kernel is employed in the nonparametric density estimator, Rudemo (1982) has shown that for any sample size, minimizing the mean integrated squared error is equivalent to minimizing the following function over positive values of σ:

$$(2T\sigma\sqrt{\pi})^{-1} + (T^2\sigma\sqrt{\pi})^{-1}$$

$$\times \sum_{t<s} \left[\exp(-\Delta_{ts}^2/4) - T\sqrt{8}\exp(-\Delta_{ts}^2/2)/(T-1)\right], \tag{13}$$

where T is the sample size and $\Delta_{ts} = (u_t - u_s)/\sigma$. In the Monte Carlo experiments that follow we minimize the sample mean integrated squared error. That is, we use the observed $\hat{u}$'s to construct σ from eq. (13) for each

trial. The resulting value of σ is then used to construct the cross-validated density estimator described in eq. (5).

4. Results

The finite sample performance of the adaptive estimator can be judged in two ways. One is to compare the efficiency of the adaptive estimator with that of its OLS or GLS counterpart, depending upon the correlation structure of the errors. This reveals the efficiency gains arising from the use of adaptive estimators when a researcher feels that the assumption of normality is not justified but the exact distribution is unknown. Another important indicator of finite sample performance is obtained by comparing the adaptive estimator with the linearized likelihood estimator. This is also constructed using the one-step procedure given in eq. (8), the difference is that the LLE uses the actual score function while the adaptive estimator uses the nonparametric estimate of the score based upon eq. (6). We impose the same trimming conditions on each, and thus can compare these two to gauge how well the nonparametric estimate of the score function is performing. Our results are based on 10,000 Monte Carlo replications of each experiment unless otherwise noted. We carry out each of these experiments with samples of size 50 and 500.

To measure the efficiency gains we calculate the root mean squared error (RMSE) directly from the sampling distributions of the estimators. We focus on the performance of the estimator of the slope coefficient, β, because under asymmetrical innovation distributions α is not adaptively estimable and the parameters of the error distribution are typically of secondary importance. For a sample of size 50 the results are given in tables 2 through 4. Each table corresponds to a different ARMA process for the errors. In addition, each table is divided into two panels, the top panel employs independent regressors while the bottom panel uses serially correlated regressors. Since the results are broadly similar across both classes of regressors, we will treat them simultaneously unless we explicitly state otherwise.

Table 2 presents the results of the experiment in which the errors are independent. The performance of the OLS estimator is roughly constant across trials showing that it is robust for the given distributions. Under normality, the OLS estimator is equivalent to the maximum likelihood estimator (MLE), and since the LLE uses it as its starting value, no additional step is necessary producing the equivalence of the LLE and the OLS estimator. As the departure from normality grows, the RMSE of the LLE typically falls indicating the potential efficiency gains. One notable exception to this is the bimodal 3 mixture where the RMSE of the LLE is substantially larger than for the bimodal 2 mixture. To understand this note that the LLE

Table 2

Root mean squared error for $\hat{\beta}$, $T = 50$.

$\varepsilon_t \sim$ Independent

	OLS	ADAPT[a]	LLE
		$x \sim$ Bernoulli	
Normal	0.2888	0.3191	0.2888
Unimodal 1	0.2876	0.3194	0.2795
Unimodal 2	0.2857	0.2766	0.2400
Unimodal 3	0.2834	0.1982	0.1619
Bimodal 1	0.2871	0.3206	0.2816
Bimodal 2	0.2865	0.1990	0.1572
Bimodal 3	0.2855	0.2702	0.2490
Lognormal 1	0.2888	0.2993 (0.2883)	0.2456
Lognormal 2	0.2827	0.2903 (0.2768)	0.2724
Lognormal 3	0.2883	0.2385 (0.2249)	0.2700
		$x \sim$ Lognormal, AR(1)	
Normal	0.1470	0.1667	0.1470
Unimodal 1	0.1466	0.1643	0.1446
Unimodal 2	0.1463	0.1480	0.1285
Unimodal 3	0.1453	0.1140	0.0964
Bimodal 1	0.1456	0.1749	0.1490
Bimodal 2	0.1453	0.1114	0.1025
Bimodal 3	0.1453	0.1091	0.1378
Lognormal 1	0.1473	0.1584 (0.1565)	0.1334
Lognormal 2	0.1473	0.1584 (0.1565)	0.1539
Lognormal 3	0.1503	0.1187 (0.1212)	0.1265

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the value in parentheses corresponds to $tr_3 = 16$.

is a one-step estimator and performs best when the likelihood function is globally quadratic. The bimodal 3 mixture is far from globally quadratic, impairing the performance of the LLE.

When comparing the adaptive estimator with the OLS estimator, the efficiency gains of the former tend to increase as the departure from normality grows. In the case of the normal and 'nearly normal' distributions (unimodal 1, bimodal 1, and lognormal 1) the adaptive estimator's RMSE is roughly 10 percent larger than the RMSE for the OLS estimator. For these slight departures from normality any potential gains available from estimating the score function are outweighed by the costs resulting from the imprecision of the score estimators.

For the other symmetric distributions the adaptive estimator provides substantial efficiency gains. With the exception of the unimodal 2 mixture, for

which the adaptive and OLS estimators are comparable, the RMSE of the adaptive estimator is on average 30 percent smaller than its OLS counterpart for a sample with only 50 observations. For the remaining two lognormal densities we find that only the lognormal 3 represents a departure from normality that is substantial enough to provide efficiency gains and here the RMSE of the adaptive estimator shrinks by 25 percent.

The extreme peaks of several of the bimodal and lognormal densities causes the ratio of $\dot{f}$ to f to far exceed that for the normal distribution in neighborhoods of these peaks. Since many of the observations fall in this range, excessive trimming may result from our initial choice of tr_3. To investigate this, we selected different values for tr_3, and found that they typically impaired the performance of the adaptive estimators for the bimodal densities. Under the lognormal densities the results were less clear. Surprisingly, the gains from relaxing the trimming parameter do not necessarily increase with the departure from normality and they carry with them a reduction in the accuracy of the standard errors (a point we will address below in our discussion of the empirical size of the tests). We have reported both results for comparison purposes.

In comparing the adaptive estimator with its LLE counterpart we are able to measure the extent to which the nonparametric estimator of the score function captures the potential efficiency gains. As described above, under the normal and 'nearly normal' distributions the adaptive estimators achieve none of the gain. For most of the other symmetric mixtures, the adaptive estimators capture between 20 and 80 percent of the possible efficiency gains. The two exceptions are the unimodal 2 and bimodal 3 mixtures, both in the lower panel. For the unimodal 2 mixture the adaptive estimator has a slightly larger RMSE than the OLS estimator, indicating that with only 50 observations the density is too similar to a Gaussian to provide any efficiency gains from estimation. When the underlying density is the bimodal 3, the adaptive estimator outperforms the LLE. This seems anomalous. While the two estimators are asymptotically equivalent, the LLE should dominate the adaptive estimator in finite samples. To understand this finding, observe that the two steep peaks of the bimodal 3 mixture make it difficult for a one-step estimator to maximize the likelihood function with a small amount of data. The adaptive estimator relies upon a smoothed version of the empirical density, and this smoothing improves the performance of a one-step estimator.

Concentrating upon the case in which $tr_3 = 8$, we have mixed results when the density is lognormal. As alluded to above, for the lognormal 1 and lognormal 2 densities, OLS outperforms the adaptive estimator for samples of size 50 and none of the efficiency gains are realized. For the lognormal 3, the adaptive estimator has a smaller RMSE than the LLE. In this case, as in

Table 3

Root mean squared error for $\hat{\beta}$, $T = 50$.

$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.3311	0.2612	0.3271	0.2663
Unimodal 1	0.3306	0.2600	0.3143	0.2610
Unimodal 2	0.3283	0.2565	0.2795	0.2243
Unimodal 3	0.3233	0.2496	0.2358	0.1962
Bimodal 1	0.3353	0.2546	0.3116	0.2663
Bimodal 2	0.3317	0.2528	0.2298	0.2232
Bimodal 3	0.3181	0.2538	0.2156	0.2975
Lognormal 1	0.3239	0.2629	0.2917 (0.2888)	0.2371
Lognormal 2	0.3237	0.2629	0.2938 (0.2871)	0.2538
Lognormal 3	0.3118	0.2581	0.2269 (0.2326)	0.2718
		$x \sim$ Lognormal, AR(1)		
Normal	0.2229	0.1992	0.2456	0.2057
Unimodal 1	0.2202	0.1992	0.2364	0.2012
Unimodal 2	0.2170	0.1997	0.2124	0.1803
Unimodal 3	0.2149	0.2012	0.1697	0.1523
Bimodal 1	0.2256	0.1897	0.2443	0.2161
Bimodal 2	0.2278	0.1879	0.1954	0.1847
Bimodal 3	0.2283	0.1884	0.1876	0.2198
Lognormal 1	0.2258	0.2022	0.2252 (0.2128)	0.1985
Lognormal 2	0.2258	0.2022	0.2223 (0.2227)	0.2207
Lognormal 3	0.2313	0.1960	0.1884 (0.2020)	0.2066

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the value in parentheses corresponds to $tr_3 = 16$.

the case of the bimodal 3 mixture, the adaptive estimator uses a smoothed likelihood function. In addition, between 5 and 10 percent of the observations are trimmed from the LLE because they have a negative value.

The results from the experiments in which ε follows an AR(1) are presented in table 3. Once again we see that OLS provides an estimator that is robust to the underlying distribution, but it is dominated by the feasible generalized least squares estimators (GLS) that incorporate an estimator of the nonscalar covariance matrix of the errors. A comparison of the GLS estimator with the LLE is consistent with the table 2 findings regarding the OLS estimator and the LLE. Introducing serial correlation eliminates the equivalence of the least squares estimator and the LLE under normality, since the starting point for the LLE (the OLS estimator) is no longer equivalent to the MLE. Once again the shape of the bimodal 3, lognormal 2, and lognormal 3 densities causes the LLE to behave rather poorly.

With the introduction of an AR(1) error process we see that the use of the OLS estimator in the first stage has substantial efficiency effects for the adaptive estimator. Its efficiency loss relative to the GLS estimator for the normal and nearly normal densities has jumped to an average increase in the RMSE of 20 percent. This is only slightly less than the efficiency loss that results when the serial dependence in ε is ignored and OLS is used. The other symmetric distributions provide similar results, an average reduction in the RMSE of only 8 percent for the adaptive estimator, while for the unimodal 2 mixture the adaptive estimator has an RMSE that is roughly 8 percent larger. The remaining two lognormal densities present the same pattern with the efficiency loss growing to 10 percent for the lognormal 3.

The poor performance of the adaptive estimator just described reduces the range of experiments over which it achieves efficiency gains. For the cases in which we can compare the adaptive estimator with the LLE, we find that the estimator of the density is not working quite as well in the AR(1) experiment. For the unimodal 3 mixture the semiparametric estimator captures roughly 45 percent of the efficiency gains indicated by the LLE, as opposed to the 65 percent captured in the previous experiment. For the bimodal 2 mixture the captured gains are reduced from 75 to roughly 15 percent in moving from table 2 to table 3. Once again for both the bimodal 3 and lognormal 3 densities the adaptive estimator performs substantially better than the LLE owing to its smoothed likelihood function.

The fourth table presents the root mean squared errors associated with the experiments in which the errors follow an MA(1) specification. Here the results differ dramatically from those presented in table 3. In moving from an autoregressive error process to a moving average error process we have reduced the efficiency of the GLS estimators relative to their OLS counterparts. This is reflective of the difficulty that arises in any moving average estimation problem. Unlike an autoregression, the regressors in a moving average are unobservable and the need to determine their value increases the variance of the error parameter estimators. This effect on the GLS estimators is so pronounced that in the lower panel, representing the situation in which β can be more precisely estimated, the RMSE of the OLS estimators is 10 percent smaller than the GLS estimator on average.

In contrast to the results from table 3, the use of the OLS estimator in the first stage of forming the adaptive estimator no longer results in substantial efficiency losses. In comparing the semiparametric and least squares methods we find that even for the normal and nearly normal densities the adaptive estimator has a smaller RMSE than the GLS. For the other symmetric mixtures the adaptive estimator outperforms the GLS estimator in all cases with an average reduction in the RMSE of 20 percent. The remaining lognormal densities provide similar results; the adaptive estimator has a smaller RMSE in all cases with an average reduction of 15 percent.

Table 4

Root mean squared error for $\hat{\beta}$, $T = 50$.

$\epsilon_t = u_t + 0.5u_{t-1}$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.2993	0.2978	0.3110	0.2617
Unimodal 1	0.3158	0.2960	0.3105	0.2742
Unimodal 2	0.3143	0.2876	0.2811	0.2496
Unimodal 3	0.3135	0.2735	0.2261	0.2027
Bimodal 1	0.3028	0.3220	0.3069	0.2659
Bimodal 2	0.3195	0.3105	0.2480	0.2364
Bimodal 3	0.3051	0.2970	0.2398	0.2851
Lognormal 1	0.3217	0.3032	0.2990(0.2883)	0.2587
Lognormal 2	0.3192	0.3038	0.3030(0.2934)	0.2769
Lognormal 3	0.3017	0.2843	0.2476(0.2550)	0.2678
		$x \sim$ Lognormal, AR(1)		
Normal	0.1975	0.2207	0.2107	0.1905
Unimodal 1	0.1934	0.2177	0.2252	0.1838
Unimodal 2	0.1931	0.2198	0.2015	0.1786
Unimodal 3	0.2037	0.2133	0.1706	0.1449
Bimodal 1	0.1992	0.2269	0.2256	0.1980
Bimodal 2	0.1942	0.2149	0.1797	0.1603
Bimodal 3	0.1931	0.2186	0.1670	0.1814
Lognormal 1	0.1860	0.2280	0.2045(0.2012)	0.1729
Lognormal 2	0.1960	0.2283	0.2059(0.1934)	0.1975
Lognormal 3	0.2081	0.2179	0.1631(0.1646)	0.1903

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the value in parentheses corresponds to $tr_3 = 16$.

To investigate the sensitivity of our results we carry out the same sequence of experiments with a sample size of 500. Table 8 corresponds to the design with independent errors. Since the above results indicate that altering the trimming parameter does more harm than good, we focus on $tr_3 = 8$. Once again the performance of the OLS estimator is roughly constant across distributions, and the increase in the number of observations leads directly to a reduction in the corresponding RMSE by a factor of 3 in the upper panel and by a factor of 5 in the lower panel. In constructing the adaptive estimators, we have employed the smoothing parameter values that were chosen from our experiments with 50 observations. Using the entire sample to form the smoothing parameter affects the RMSE only slightly but increases the empirical test size dramatically. We address this point at greater length in the discussion on empirical test sizes given below.

Table 5

Empirical size of a nominal 5 percent test, $T = 50$.

$\varepsilon_t \sim$ Independent

	OLS	ADAPT[a]	LLE
		$x \sim$ Bernoulli	
Normal	0.0556	0.2530	0.0834
Unimodal 1	0.0546	0.2710	0.0781
Unimodal 2	0.0557	0.2760	0.0576
Unimodal 3	0.0432	0.2560	0.0262
Bimodal 1	0.0627	0.2200	0.0793
Bimodal 2	0.0601	0.2890	0.0247
Bimodal 3	0.0598	0.3660	0.0620
Lognormal 1	0.0538	0.2920(0.3540)	0.0603
Lognormal 2	0.0538	0.2540(0.3160)	0.0742
Lognormal 3	0.0560	0.3130(0.4420)	0.0729
		$x \sim$ Lognormal, AR(1)	
Normal	0.0572	0.2390	0.0216
Unimodal 1	0.0582	0.2250	0.0209
Unimodal 2	0.0593	0.2710	0.0165
Unimodal 3	0.0618	0.2670	0.0093
Bimodal 1	0.0581	0.2380	0.0222
Bimodal 2	0.0544	0.3110	0.0105
Bimodal 3	0.0544	0.2100	0.0190
Lognormal 1	0.0579	0.2730(0.3330)	0.0178
Lognormal 2	0.0579	0.2730(0.3330)	0.0237
Lognormal 3	0.0562	0.2890(0.4660)	0.0160

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the numbers in parentheses correspond to $tr_3 = 16$.

In comparing the adaptive estimator with OLS we still find efficiency losses of nearly 10 percent when the errors are normally distributed. For the 'nearly normal' distributions, the increase in the sample size reduces, and in one case eliminates, the efficiency losses suffered when estimating the score function nonparametrically. The most striking gains are for the lognormal 1, where the adaptive estimator has an RMSE that is between 10 and 20 percent smaller than its OLS counterpart. The two estimators are now roughly equivalent for the bimodal 1, while under the unimodal 1, the closest density to the normal, the adaptive estimator still suffers losses of slightly less than 10 percent.

The other symmetric distributions reveal similar gains. For the remaining bimodal mixtures, the adaptive estimator has an RMSE that is on average 2.5 times smaller than its OLS counterpart. For the unimodal mixtures the

Table 6

Empirical size of a nominal 5 percent test, $T = 50$.

$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.0588	0.0400	0.3100	0.0530
Unimodal 1	0.0670	0.0300	0.2940	0.0580
Unimodal 2	0.0640	0.0400	0.2980	0.0600
Unimodal 3	0.0480	0.0400	0.3310	0.0980
Bimodal 1	0.0580	0.0400	0.2770	0.0420
Bimodal 2	0.0610	0.0500	0.3520	0.0800
Bimodal 3	0.0470	0.0600	0.3260	0.1170
Lognormal 1	0.0610	0.0400	0.3150(0.3620)	0.0220
Lognormal 2	0.0610	0.0500	0.2940(0.3460)	0.0860
Lognormal 3	0.0420	0.0300	0.2770(0.4520)	0.1590
		$x \sim$ Lognormal, AR(1)		
Normal	0.1720	0.0400	0.2920	0.0710
Unimodal 1	0.1720	0.0700	0.3120	0.0830
Unimodal 2	0.1600	0.0700	0.3360	0.1070
Unimodal 3	0.1530	0.0300	0.3770	0.1680
Bimodal 1	0.1860	0.0900	0.2790	0.0710
Bimodal 2	0.1860	0.0500	0.4210	0.0730
Bimodal 3	0.1840	0.0600	0.3380	0.1100
Lognormal 1	0.1790	0.1000	0.3130(0.3660)	0.0700
Lognormal 2	0.1800	0.0700	0.3080(0.3860)	0.0820
Lognormal 3	0.1710	0.0400	0.3630(0.5470)	0.1210

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the numbers in parentheses correspond to $tr_3 = 16$.

adaptive estimator outperforms OLS in each case, reducing the RMSE by roughly 10 percent for the second unimodal mixture and by a factor of 2 under the unimodal 3 density. With a sample of size 500, the adaptive estimator is able to exploit the departures from normality contained in each of the remaining lognormal densities shrinking the RMSE by more than 60 percent.

To measure the effects of the sample size on the nonparametric density estimator we focus on the performance of the adaptive estimator relative to the LLE. With the exception of the lognormal 1, the adaptive estimator is unable to capture any of the efficiency gains for the normal and 'nearly normal' densities. For the remaining densities, the captured gains increase by at least 20 percentage points when the sample size grows. Once again the adaptive estimator outperforms the LLE for the third density in both the bimodal and lognormal classes indicating that the smoothing induced by the kernel estimator improves the performance when only one step is taken.

Table 7

Empirical size of a nominal 5 percent test, $T = 50$.

$\varepsilon_t = u_t + 0.5u_{t-1}$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.0430	0.1300	0.3190	0.1060
Unimodal 1	0.0560	0.1300	0.3200	0.1410
Unimodal 2	0.0500	0.1200	0.3150	0.1330
Unimodal 3	0.0490	0.0700	0.3360	0.1370
Bimodal 1	0.0540	0.1200	0.3040	0.1400
Bimodal 2	0.0610	0.1200	0.3340	0.1310
Bimodal 3	0.0520	0.1600	0.3470	0.1960
Lognormal 1	0.0660	0.1500	0.3940(0.3860)	0.1200
Lognormal 2	0.0600	0.1000	0.3440(0.3810)	0.0860
Lognormal 3	0.0600	0.1100	0.3480(0.5190)	0.1860
		$x \sim$ Lognormal, AR(1)		
Normal	0.1110	0.1000	0.2860	0.1100
Unimodal 1	0.1180	0.1400	0.3150	0.1080
Unimodal 2	0.1230	0.1200	0.3330	0.1130
Unimodal 3	0.1210	0.0700	0.3760	0.1240
Bimodal 1	0.1300	0.1300	0.2760	0.1050
Bimodal 2	0.1080	0.1100	0.3820	0.1550
Bimodal 3	0.1400	0.1800	0.3480	0.1730
Lognormal 1	0.1200	0.1400	0.3360(0.3720)	0.1080
Lognormal 2	0.1330	0.1100	0.3110(0.3490)	0.1590
Lognormal 3	0.1130	0.1300	0.3490(0.4980)	0.1840

[a]Based on 1000 Monte Carlo simulations. For the lognormal densities, the numbers in parentheses correspond to $tr_3 = 16$.

Table 9 presents the simulation with an AR(1) error when the sample size is 500. In comparing the GLS estimators with the adaptive estimators, the efficiency losses for the normal and nearly normal densities now average less than 5 percent across both panels. In one case, the lognormal 1 in the upper panel, the adaptive estimator's RMSE is nearly 20 percent smaller than its GLS counterpart. For the remaining symmetric distributions the adaptive estimator has an RMSE that is on average 75 percent less than its GLS counterpart, and for three of the eight simulations the GLS RMSE is more than twice as large as its adaptive counterpart. With the increase in the sample size substantial efficiency gains are realized for all but one of the remaining lognormal simulations, and in one case again the GLS RMSE is double the size of the adaptive RMSE.

In comparing the adaptive estimator with the LLE, the larger sample size has reduced to 7 the number of simulations in which the adaptive estimator

Table 8

Root mean squared error for $\hat{\beta}$, $T = 500$.

$\epsilon_t \sim$ Independent

	OLS	ADAPT[a]	LLE
	$x \sim$ Bernoulli		
Normal	0.0887	0.0972	0.0887
Unimodal 1	0.0900	0.0939	0.0882
Unimodal 2	0.0898	0.0824	0.0735
Unimodal 3	0.0896	0.0389	0.0344
Bimodal 1	0.0901	0.0918	0.0858
Bimodal 2	0.0882	0.0323	0.0303
Bimodal 3	0.0900	0.0456	0.0465
Lognormal 1	0.0902	0.0750	0.0784
Lognormal 2	0.0900	0.0787	0.0729
Lognormal 3	0.0901	0.0437	0.0976
	$x \sim$ Lognormal, AR(1)		
Normal	0.0342	0.0372	0.0342
Unimodal 1	0.0340	0.0360	0.0335
Unimodal 2	0.0344	0.0322	0.0288
Unimodal 3	0.0343	0.0172	0.0160
Bimodal 1	0.0344	0.0344	0.0331
Bimodal 2	0.0337	0.0149	0.0157
Bimodal 3	0.0344	0.0117	0.0264
Lognormal 1	0.0339	0.0309	0.0298
Lognormal 2	0.0339	0.0310	0.0295
Lognormal 3	0.0345	0.0213	0.0430

[a]Based on 1000 Monte Carlo simulations.

captures none of the gains achieved by the LLE. For the second and third unimodal densities the realized efficiency gains increase by 20 percent. For the corresponding bimodal mixtures the larger sample size allows the adaptive estimator to dominate its one-step counterpart everywhere. As alluded to earlier this results from the smoothed likelihood function created from the nonparametric kernel estimator which presents an easier one-step minimization problem than the true likelihood when the latter is not well approximated by a quadratic function. This carries over to the final two lognormal densities where efficiency gains are now realized in three of the four experiments, in two of which the adaptive estimator outperforms the LLE.

The MA(1) experiment with the larger sample size is presented in table 10. Efficiency gains are now realized for every simulation in both panels. Gains average nearly 15 percent for the normal and nearly normal densities and 70 percent for the additional two lognormal densities. The remaining symmetric distributions are characterized by GLS RMSE's that on average are more

Table 9

Root mean squared error for $\hat{\beta}$, $T = 500$.

$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.1031	0.0775	0.0876	0.0801
Unimodal 1	0.1050	0.0819	0.0878	0.0799
Unimodal 2	0.1044	0.0800	0.0746	0.0666
Unimodal 3	0.1049	0.0812	0.0364	0.0340
Bimodal 1	0.1023	0.0768	0.0827	0.0773
Bimodal 2	0.1043	0.0806	0.0305	0.0328
Bimodal 3	0.1038	0.0768	0.0421	0.0694
Lognormal 1	0.1033	0.0825	0.0692	0.0715
Lognormal 2	0.1021	0.0849	0.0738	0.0669
Lognormal 3	0.1032	0.0866	0.0371	0.1176
		$x \sim$ Lognormal, AR(1)		
Normal	0.0565	0.0490	0.0498	0.0470
Unimodal 1	0.0557	0.0447	0.0517	0.0456
Unimodal 2	0.0572	0.0447	0.0431	0.0407
Unimodal 3	0.0573	0.0436	0.0302	0.0292
Bimodal 1	0.0565	0.0469	0.0492	0.0456
Bimodal 2	0.0578	0.0458	0.0302	0.0346
Bimodal 3	0.0574	0.0447	0.0222	0.0473
Lognormal 1	0.0566	0.0458	0.0450	0.0436
Lognormal 2	0.0566	0.0469	0.0471	0.0464
Lognormal 3	0.0569	0.0480	0.0361	0.0568

[a]Based on 1000 Monte Carlo simulations.

than twice as large as the adaptive RMSE's. The adaptive estimator now captures at least 60 percent of the available gains for all the nonnormal densities, and in half of the cases it outperforms the LLE.

The above discussion pertains exclusively to the RMSE constructed from the sampling distributions of the estimators. While providing much useful information on the accuracy of the estimators, it certainly is not the only information of interest to an applied researcher. In reporting a confidence interval for a given estimation problem one relies upon the accuracy of the standard error as well. Since estimated standard errors are known to underestimate the true standard error in many cases, such as GLS estimation, one wonders to what degree this problem characterizes adaptive estimators.

To address this concern, we construct a t-test of the null hypothesis that β equals 1 against a two-sided alternative. Our test is based upon a nominal size of 5 percent; in tables 5 through 7 we report the empirical size of the tests for a sample size of 50, while in tables 11 through 13 we use 500 observations.

Table 10

Root mean squared error for $\hat{\beta}$, $T = 500$.

$\varepsilon_t = u_t + 0.5u_{t-1}$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.1021	0.0879	0.0858	0.0796
Unimodal 1	0.0999	0.0934	0.0838	0.0768
Unimodal 2	0.1007	0.0889	0.0730	0.0657
Unimodal 3	0.0996	0.0852	0.0345	0.0353
Bimodal 1	0.0994	0.0959	0.0794	0.0752
Bimodal 2	0.1004	0.0937	0.0311	0.0356
Bimodal 3	0.1003	0.0876	0.0345	0.0720
Lognormal 1	0.0999	0.0891	0.0692	0.0706
Lognormal 2	0.1010	0.0911	0.0710	0.0661
Lognormal 3	0.1009	0.0893	0.0372	0.1021
		$x \sim$ Lognormal, AR(1)		
Normal	0.0475	0.0526	0.0458	0.0435
Unimodal 1	0.0473	0.0497	0.0454	0.0425
Unimodal 2	0.0472	0.0479	0.0374	0.0363
Unimodal 3	0.0478	0.0463	0.0257	0.0238
Bimodal 1	0.0480	0.0543	0.0462	0.0422
Bimodal 2	0.0480	0.0480	0.0231	0.0270
Bimodal 3	0.0475	0.0478	0.0184	0.0387
Lognormal 1	0.0476	0.0519	0.0402	0.0392
Lognormal 2	0.0474	0.0526	0.0398	0.0400
Lognormal 3	0.0476	0.0520	0.0285	0.0503

[a]Based on 1000 Monte Carlo simulations.

The fifth table lists the results for the experiments with an independent error process. The OLS estimator again shows it is robust to a wide variety of distributional misspecifications with an empirical size that is roughly 5 percent in all cases. The LLE is more erratic, in the upper panel its empirical size ranges from 3 to 8 percent while in the lower panel the size reflects a conservative test. One factor is that the estimated covariance matrix for the LLE, formed from the outer product of the sample scores, is fragile for samples of 50 observations.

The adaptive estimator has an empirical size that averages nearly 30 percent across both panels. The estimated covariance matrix for this estimator is the same as the outer product used for the LLE with the exception that the estimated density is used in place of the actual density. A comparison of the two reveals the degree of bias introduced into the standard errors through the use of a nonparametric component. The large increase in the empirical size of the adaptive estimator relative to the LLE indicates that the

limiting distribution provides a poor guide to the distribution, based upon only 50 observations. For the lognormal densities we report the size when both values of the trimming parameter are used. Decreasing the amount of trimming introduces more bias, while reducing the variance in the kernel estimator, and as the departure from normality grows this leads to a large increase in the bias of the standard errors.

Table 6 lists the empirical sizes for the experiments in which ε follows an AR(1). For the Bernoulli regressor case the OLS estimator continues to have approximately the right size. When we move away from this case to the lower panel in which the regressor is continuous we have the standard result that the uncorrected OLS standard errors are misleading and the empirical size exceeds 15 percent. For the GLS estimators we find the test is slightly conservative in the upper panel, while the bias of the GLS standard errors is slightly less than most of their OLS counterparts in the lower panel. Once again the performance of the LLE is highly erratic and its empirical size increases substantially as the departure from normality grows. The true size is especially sensitive to densities for which a one-step estimator may not provide a good approximation such as the bimodal 3. The results for the adaptive estimator are fairly comparable to those noted above, an average size of 30 percent that is roughly constant across densities. While indicating the problems with the first-order asymptotic expansion that we have outlined above, the results for the third members of both the bimodal and lognormal families indicate the smoothed kernel density estimator can reduce the bias in the adaptive estimator and its standard errors for small samples.

The results of the MA(1) experiment are found in table 7. All of the findings are broadly consistent with the AR(1) experiment. Both least squares estimators exhibit a significant downward bias in their standard errors in the lower panel with empirical sizes between 10 and 20 percent. The LLE suffers from potential bias throughout and the degree of bias grows with the departure from normality. The adaptive estimator has an empirical test size of more than 30 percent that is relatively constant across densities. The smoothed estimator of the density reduces the true test size substantially below that of the LLE for the densities with the most significant departures from normality.

To see if this problem is principally due to the small samples size we have reported the results for $T = 500$ in tables 11–13. For the design in which the errors are independent, reported in table 11, we see that the empirical size has been greatly reduced. While most of the values range between 4 and 8 percent, the densities representing the most extreme departures from normality are characterized by empirical sizes in excess of 10 percent.

This pattern is repeated in tables 12 and 13. For the AR(1) experiment the size of the adaptive estimator is on average slightly above 10 percent. This is roughly comparable to the average LLE size and represents a significant improvement over the results reported in table 6, where the size for the

Table 11

Empirical size of a nominal 5 percent test, $T = 500$.

$\varepsilon_t \sim$ Independent

	OLS	ADAPT[a]	LLE
		$x \sim$ Bernoulli	
Normal	0.0469	0.0730	0.0468
Unimodal 1	0.0522	0.0860	0.0493
Unimodal 2	0.0520	0.1170	0.0476
Unimodal 3	0.0523	0.0930	0.0632
Bimodal 1	0.0528	0.0640	0.0495
Bimodal 2	0.0492	0.0440	0.0664
Bimodal 3	0.0541	0.1150	0.0980
Lognormal 1	0.0526	0.0410	0.0487
Lognormal 2	0.0509	0.0570	0.0533
Lognormal 3	0.0477	0.1130	0.0782
		$x \sim$ Lognormal, AR(1)	
Normal	0.0516	0.0610	0.0483
Unimodal 1	0.0508	0.0880	0.0495
Unimodal 2	0.0557	0.1220	0.0524
Unimodal 3	0.0535	0.1520	0.1380
Bimodal 1	0.0533	0.0520	0.0443
Bimodal 2	0.0452	0.1150	0.0772
Bimodal 3	0.0519	0.0530	0.1113
Lognormal 1	0.0480	0.0640	0.0193
Lognormal 2	0.0497	0.0640	0.0578
Lognormal 3	0.0494	0.1630	0.1461

[a]Based on 1000 Monte Carlo simulations.

adaptive estimator was roughly three times that for the LLE. Similar results are shown in table 13 where again the test size for the adaptive estimator is roughly equivalent with that of the LLE. This increase in the sample size also improves the precision with which the error parameter is estimated. This leads directly to a reduction in the test size of the GLS estimator as can be seen most noticeably by comparing tables 7 and 13.

While the larger sample leads to a decrease in the empirical test size of the adaptive estimator, the size is still related to the underlying density. In a number of cases, the empirical size increases with the degree of the departure from normality. As can be seen in table 14, these are the densities typically associated with the smallest value of σ.

This finding relates directly to the choice of the smoothing parameter discussed earlier. When the full sample is used to construct a cross-validated estimator of the density, the smoothing parameters chosen are typically much smaller than when a sample of 50 observations is used. While this accords

Table 12

Empirical size of a nominal 5 percent test, $T = 500$.

$\varepsilon_t = 0.5\varepsilon_{t-1} + u_t$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.0518	0.0400	0.0890	0.0477
Unimodal 1	0.0528	0.0470	0.0970	0.0545
Unimodal 2	0.0533	0.0450	0.1190	0.0499
Unimodal 3	0.0541	0.0550	0.1050	0.0938
Bimodal 1	0.0504	0.0410	0.0710	0.0526
Bimodal 2	0.0574	0.0510	0.0580	0.1222
Bimodal 3	0.0544	0.0440	0.1510	0.1389
Lognormal 1	0.0504	0.0550	0.0470	0.0200
Lognormal 2	0.0488	0.0550	0.0870	0.0596
Lognormal 3	0.0495	0.0630	0.0930	0.1202
		$x \sim$ Lognormal, AR(1)		
Normal	0.1781	0.0600	0.0840	0.0541
Unimodal 1	0.1664	0.0500	0.1010	0.0603
Unimodal 2	0.1808	0.0480	0.1380	0.0749
Unimodal 3	0.1760	0.0480	0.2380	0.1605
Bimodal 1	0.1734	0.0540	0.0670	0.0565
Bimodal 2	0.1852	0.0470	0.2240	0.1182
Bimodal 3	0.1815	0.0480	0.1220	0.1512
Lognormal 1	0.1725	0.0490	0.0890	0.0397
Lognormal 2	0.1807	0.0610	0.0940	0.0980
Lognormal 3	0.1686	0.0490	0.2460	0.1386

[a]Based on 1000 Monte Carlo simulations.

with the requirement that σ shrink as T grows larger, it leads to values of σ that are too small for a reliable estimator of the derivative of the density.

An alternative method for selecting σ is the bootstrap algorithm employed by Hsieh and Manski (1987). While their experiments don't match ours exactly, there is some overlap for $T = 50$. Since the results are relatively insensitive to both the choice of the regressors and correlation structure of the errors, table 14 presents the smoothing parameter values only for the model in which both the regressors and the errors are serially uncorrelated. As mentioned above, when the departure from normality grows, the optimal size of the smoothing parameter shrinks. The nonparametric kernel is simply reacting to the clustering of these densities, as the concentration of points increases outlying observations are given smaller weights. For the four densities that overlap, we have presented the optimal values of the smoothing parameter selected by Hsieh and Manski. Using the mean square error criterion function, the smoothing parameter selected by the data in three out

Table 13

Empirical size of a nominal 5 percent test, $T = 500$.

$\varepsilon_t = u_t + 0.5u_{t-1}$

	OLS	GLS[a]	ADAPT[a]	LLE
		$x \sim$ Bernoulli		
Normal	0.0578	0.0660	0.0880	0.0621
Unimodal 1	0.0500	0.0900	0.0990	0.0522
Unimodal 2	0.0545	0.1130	0.1250	0.0577
Unimodal 3	0.0483	0.0480	0.0950	0.1226
Bimodal 1	0.0502	0.0950	0.0610	0.0584
Bimodal 2	0.0511	0.0580	0.0650	0.1280
Bimodal 3	0.0512	0.0720	0.1520	0.1939
Lognormal 1	0.0484	0.1000	0.0700	0.0279
Lognormal 2	0.0523	0.0450	0.0810	0.0765
Lognormal 3	0.0509	0.0830	0.1070	0.1663
		$x \sim$ Lognormal, AR(1)		
Normal	0.1143	0.0750	0.0780	0.0577
Unimodal 1	0.1133	0.0410	0.0910	0.0573
Unimodal 2	0.1135	0.0920	0.1150	0.0661
Unimodal 3	0.1148	0.0630	0.1770	0.1148
Bimodal 1	0.1235	0.0580	0.0640	0.0572
Bimodal 2	0.1234	0.0940	0.1870	0.0755
Bimodal 3	0.1189	0.1350	0.1110	0.1131
Lognormal 1	0.1181	0.0640	0.0630	0.0328
Lognormal 2	0.1163	0.0510	0.0670	0.0886
Lognormal 3	0.1097	0.0920	0.1730	0.1230

[a]Based on 1000 Monte Carlo simulations.

Table 14

Optimal values of the smoothing parameter.

	MSE criterion	HM bootstrap
Normal	0.23	0.05
Unimodal 1	0.22	
Unimodal 2	0.18	
Unimodal 3	0.09	0.10
Bimodal 1	0.24	
Bimodal 2	0.10	0.10
Bimodal 3	0.04	
Lognormal 1	0.22	
Lognormal 2	0.20	
Lognormal 3	0.07	0.10

of the four cases is equivalent with the value chosen by their bootstrap design. In the remaining case, the normal, the distribution of the estimators is quite robust to the choice of the smoothing parameter and the optimizing function is extremely flat. The equivalence of these two techniques indicates that bootstrapping the nonparametric density estimator may lead to values of the smoothing parameter that are too small for reliable score function estimation. A more promising approach is to apply either the bootstrap or cross-validation methods directly to the nonparametric estimator of the score function. In this way one may avoid the selection of values of σ that cause the estimator of the derivative of f to be unstable.

5. Conclusion

The finite sample results presented above are quite striking, indicating that a large sample size is not needed to generate substantial efficiency gains. Across a range of distributions with significant departures from normality the adaptive estimators produce root mean squared errors that are on average 10 to 30 percent smaller than their OLS or GLS counterparts for a sample of only 50 observations. When the sample is increased to 500 observations the efficiency gains are even more pronounced. Yet for distributions that exhibit only slight departures from normality the estimators can suffer efficiency losses of 10 to 15 percent. This warns against the use of adaptive procedures unless nonnormality is strongly suspected.

Further, the empirical size of the adaptive estimators is quite poor for very small samples, in our experiments nearly 30 percent for a nominal size of 5 percent. Increasing the sample size brings about rapid gains, reducing the empirical size to 10 percent, roughly equivalent to the linearized likelihood estimator it is trying to emulate.

The results presented above rely on parsimonious representations of the error process. However, we believe our findings are applicable to a wide range of potential ARMA models. Although the adaptive estimators failed to capture efficiency gains for the first-order autoregressive process, these results may be misleading. For this special case the results are extremely sensitive to the choice of the relatively inefficient OLS estimator as the starting value. A more relevant example is provided by the first-order moving average process. It is merely the simplest way to characterize the difficulties which arise in estimating more general ARMA models. For this case the adaptive estimators exhibit efficiency gains of nearly the same magnitude as in the case where the errors are independent. Therefore, adaptive estimators are potentially applicable in a wide range of circumstances.

While this paper answers several important questions, it also raises others. Future work should address the problem of selecting the smoothing parameter, perhaps by examining the possibility of bootstrapping the score function directly. There is also room for improvement in the nonparametric estimator

of the density. A technique that incorporates a variable bandwidth, such as nearest neighbor estimation, may be helpful here. Using it one would not need to specify trimming parameters and could avoid any problem of excess trimming.

We hope that this work will provide researchers with a road map to these relatively uncharted regions of estimation. While adaptive estimators present a powerful alternative to robust methods of estimation, they should not be applied blindly. As this work shows, they are most appropriate in settings in which there is reason to believe that the errors exhibit substantial departures from normality. Under these circumstances, when the problem of precise estimation is interesting adaptive procedures make it attainable.

Appendix

The random numbers used in this study are generated in the following way. For each sample, a call is made to the internal clock time which randomizes the seed used in the Cray intrinsic function ranf. The independent, uniform [0, 1] numbers are then transformed into the draws from a normal $(0, \frac{1}{2})$ distribution using the following formulas:

$$X_i = (-\ln Z_i)^{1/2}\cos(2\pi Z_{i+1}), \qquad X_{i+1} = (-\ln Z_i)^{1/2}\sin(2\pi Z_{i+1}),$$

where $Z \sim \mathrm{U}[0, 1]$ and $X \sim \mathrm{N}(0, \frac{1}{2})$. All of the distributions we employ in this study are simple transformations of this normal distribution.

References

Bickel, P., 1982, On adaptive estimation, Annals of Statistics 10, 647–671.
Hsieh, D. and C. Manski, 1987, Monte Carlo evidence on adaptive maximum likelihood estimation of a regression, Annals of Statistics 15, 541–551.
Manski, C., 1984, Adaptive estimation of non-linear regression models, Econometric Reviews 3, 145–194.
Newey, W., 1990, Semiparametric efficiency bounds, Journal of Applied Econometrics 5, 99–135.
Rudemo, M., 1982, Empirical choice of histograms and kernal density estimators, Scandinavian Journal of Statistics 9, 65–98.
Steigerwald, D., 1990, Generalized adaptive estimation in econometric and financial models, Manuscript (University of California, Santa Barbara, CA).
Steigerwald, D., 1992, Adaptive estimation in time series regression models, Journal of Econometrics, forthcoming.
White, H. and G. MacDonald, 1980, Some large-sample tests for nonnormality in the linear regression model, Journal of the American Statistical Association 75, 16–31.