Bias-adjusted estimation in the ARX(1) model

Broda, S; Paolella, M S; Carstensen, K

Postprint available at:
http://www.zora.uzh.ch

Posted at the Zurich Open Repository and Archive, University of Zurich.
http://www.zora.uzh.ch

Originally published at:
Computational Statistics &amp; Data Analysis 2007, 51(7):3355-3367.
Bias-adjusted estimation in the ARX(1) model

Abstract

A new point estimator for the AR(1) coefficient in the linear regression model with arbitrary exogenous regressors and stationary AR(1) disturbances is developed. Its construction parallels that of the median--unbiased estimator, but uses the mode as a measure of central tendency. The mean--adjusted estimator is also considered, and saddlepoint approximations are used to lower the computational burden of all the estimators. Large--scale simulation studies for assessing their small--sample properties are conducted. Their relative performance depends almost exclusively on the value of the autoregressive parameter, with the new estimator dominating over a large part of the parameter space.
Bias–Adjusted Estimation in the ARX(1) Model

Simon Broda, a  Kai Carstensen b  Marc S. Paolella a *

a Swiss Banking Institute, University of Zurich, Switzerland
b Institute for World Economics, Kiel, Germany

June 2006

Abstract

A new point estimator for the AR(1) coefficient in the linear regression model with arbitrary exogenous regressors and stationary AR(1) disturbances is developed. Its construction parallels that of the median–unbiased estimator, but uses the mode as a measure of central tendency. The mean–adjusted estimator is also considered, and saddlepoint approximations are used to lower the computational burden of all the estimators. Large–scale simulation studies for assessing their small–sample properties are conducted. Their relative performance depends almost exclusively on the value of the autoregressive parameter, with the new estimator dominating over a large part of the parameter space.

Keywords: Autoregression, Bias Correction, Saddlepoint Approximation

*Corresponding author. Swiss Banking Institute, University of Zurich, Plattenstrasse 14, CH–8032, Zurich, Switzerland. E-mail: paolella@isb.unizh.ch, Tel: +41 1 634 4584, Fax: +41 1 634 4903.
1 Introduction

Inferential procedures for the parameters of autoregressive models (with or without covariates) continue to receive a great amount of attention in the theoretical literature, with recent contributions including work on structural breaks (Breitung, 2002; Kurozumi, 2002; Saikkonen and Lütkepohl, 2002 and the references therein), innovation variance shifts (Kim, Leybourne and Newbold, 2002), methods which use point optimal tests (Shively, 2001), local-to-unity arguments (Elliot and Stock, 2001), and bootstrap techniques (Hansen, 1999).

This paper develops a new estimator for the autoregressive coefficient in a first-order autoregression. Its derivation parallels that of the median–unbiased point estimator of Andrews (1993), but uses the mode, rather than the median, as a measure of central tendency. Both of these methods entail the relatively costly numeric evaluation of the distribution function of a ratio of quadratic forms in normal random variables, accomplished by inversion of the relevant characteristic function via, for example, the method of Imhof (1961). A way of circumventing these extensive calculations — without the restrictions associated with pre-computed tables, as provided by Andrews for his estimator — is to replace the exact evaluation of the requisite distribution function with a saddlepoint approximation, thus removing the bottleneck in the procedure so that the estimators can be calculated in about a hundredth of the time otherwise necessary. Along with these massive time savings, the accuracy of the saddlepoint approximation, or, in short, SPA, is not only high enough for practical work, but in fact results in higher accuracy than can be achieved by interpolation from pre-computed tables.

Because of the similarity in its construction, we also include in our study the mean–adjusted estimator of Tanizaki (2000), and show how it can be computed without having to resort to simulation. Owing to these numerical methods, all three bias corrected estimators can be computed fast enough to make a simulation study feasible, using a variety of data generating models. We find, somewhat surprisingly—and quite conveniently—, that the relative small–sample properties of the estimators is virtually invariant to the choice of sample size and set of regressors. Moreover, this fortuitous behavior remains (approximately) constant for a variety of nonnormal innovation distribution assumptions commonly entertained in practice. The optimal choice of estimator depends (essentially) only on the true value of the autoregressive parameter $\alpha$, but in virtually the same way for any model design and distributional assumption. For example, one of the estimators has lowest mean squared error for all $\alpha$ between about 0.7 and 1.0—a result which should be of interest when working with series with high persistence or near unit–root behavior. The ranges of $\alpha$ for which a particular estimator is optimal are all quite large, these being ($-1, -0.1$), ($-0.1, 0.7$) and ($0.7, 1$), so that even a very small amount of “prior information” on the part of the researcher can be effectively used.

The remainder of this paper is as follows. Section 2 introduces the model and relevant no-
tation. Section 3 and 4 review the median–unbiased and mean–adjusted estimators previously proposed in the literature, respectively, and show how the computational burden associated with them can be substantially reduced. Section 5 develops the mode–adjusted estimator. Section 6 details several simulation studies, from which the performance of the estimators can be assessed and recommendations made for their use. Section 7 concludes and briefly discusses some ideas for further research. The appendix provides the required formulae for the saddlepoint approximations.

2 The Model

Using notation similar to that in Andrews (1993), the model consists of an observed and a latent equation given respectively by

\[ Y_t = x_t' \beta + Y_t^\ell, \quad t = 0, \ldots, T, \tag{1} \]

and

\[ Y_t^\ell = \alpha Y_{t-1}^\ell + U_t, \quad t = 1, \ldots, T, \quad U_t \sim \text{N}(0, \sigma^2), \tag{2} \]

where \( X = [x_0, \ldots, x_T]' \) is assumed to be a full rank \((T+1) \times k\) matrix, and the initial observation \( Y_0^\ell \) follows the unconditional distribution of \( Y_t, Y_0^\ell \sim \text{N}(0, \sigma^2/(1 - \alpha^2)) \) if \( \alpha \in (-1,1) \) and an arbitrary constant if \( \alpha = 1 \). Note that a unit root \( \alpha = 1 \) implies that (1) becomes a spurious regression model. Asymptotically, in such a model, the effect of \( x_t \) on \( y_t \) is irrelevant and the parameter \( \beta \) is not meaningful to interpret. However, we work with finite samples, and our main focus is on inference for \( \alpha \). Moreover, while we do not engage in unit root testing in this paper, it is straightforward extend the methods described herein to the construction of confidence intervals, and thus, tests of the unit root hypothesis; see Andrews (1993).

We consider point estimators for \( \alpha \). Point- and interval estimates for the regression coefficients \( \beta \) can be obtained by generalized least squares, using the estimated value of \( \alpha \) in the requisite covariance matrix. However, in our experiments, the differences in inferential accuracy when using different estimators for \( \alpha \) were negligible; as such, we focus our attention on the autoregressive parameter.

The AR(1) model (1)–(2) can be estimated by least squares after combining the observable and latent equations to

\[ Y_t = Y_{t-1} \alpha + x_t' \beta - x_{t-1}' \beta \alpha + U_t, \quad t = 1, \ldots, T, \]

or, in matrix form,

\[ Y_T = Y_{T-1} \alpha + Z \gamma + U_T, \tag{3} \]
where $\gamma = [\beta', -\beta'\alpha]'$, $Y_T = [Y_1, \ldots, Y_T]'$, $Y_{T-1} = [Y_0, \ldots, Y_{T-1}]'$, $Z = [X_T, X_{T-1}]$, $X_T = [x_1, \ldots, x_T]'$, $X_{T-1} = [x_0, \ldots, x_{T-1}]'$ and $U_T = [U_1, \ldots, U_T]'$.

It should be noted that, due to the common factor restrictions on the $2k$ parameters in $\gamma$, model (3) is, in general, different from the dynamic linear model

$$Y_t = Y_{t-1}\alpha + x_t'\beta + U_t, \quad t = 1, \ldots, T.$$ 

However, as remarked by Andrews (1993, p. 146), the methodology can also be made applicable to this model by including in $Z$ a set of constructed regressors, designed to maintain the invariance property. For the specification of these regressors, see Kiviet and Phillips (1990). Neglecting these restrictions, a straightforward application of the Frisch–Waugh theorem as in Andrews (1993) shows that the ordinary least squares (OLS) estimator of $\alpha$ can be expressed as

$$\hat{\alpha}_{LS} = \frac{Y_T' MY_T}{Y_{T-1}' MY_{T-1}}, \quad (4)$$

where $M = I_T - Z(Z'Z)^{-1}Z'$. If $Z$ has reduced column rank, then it should be replaced by a full column rank matrix spanning the same column space.

As remarked by Andrews (1993, p. 146), and as is required for construction of the estimators considered herein, $\hat{\alpha}_{LS}$ is independent of $\beta$ and $\sigma^2$ (and of $Y_0$ if $\alpha = 1$) for any exogenous regressor matrix $X$. A detailed proof is given in Broda et al. (2004).

### 3 Median–unbiased Estimation

It is well known that $\hat{\alpha}_{LS}$ is downward biased, extremely so for $\alpha$ near one. While various procedures exist to partially correct for this, no operational method has so far been devised which is exactly mean–unbiased. It is, however, straightforward to construct a median–unbiased estimator, hereafter denoted $\hat{\alpha}_{Med}$, first pursued in this context by Andrews (1993). By definition, an estimator $\hat{\theta}$ is median–unbiased for $\theta$ if, for each value $\theta$ in the parameter space, $\theta$ is a median of $\hat{\theta}$. The following bias correction procedure then makes $\hat{\alpha}_{Med}$ a median–unbiased estimator: $\hat{\alpha}_{Med}$ takes that value of $\alpha$ that yields the OLS estimator to have a median equal to the OLS estimate obtained from the data. More formally, let $\text{Med}(\hat{\alpha}_{LS} \mid \alpha, X) = m(\alpha)$ denote the median function of $\hat{\alpha}_{LS}$ when $\alpha$ is the true parameter, and let $m^{-1} : (m(-1), m(1)] \rightarrow (-1, 1]$ denote its inverse. For this to be meaningful, it is required that $m(\alpha)$ be strictly increasing. As noted by Andrews, it is not apparent how the latter condition can be verified analytically; however, in the present setup, numerical evidence suggests it holds.
The median unbiased estimator $\hat{\alpha}_{\text{Med}}$ is then given by

$$
\hat{\alpha}_{\text{Med}} = \begin{cases} 
1, & \text{if } \hat{\alpha}_{\text{LS}} > m(1), \\
m^{-1}(\hat{\alpha}_{\text{LS}}), & \text{if } m(-1) < \hat{\alpha}_{\text{LS}} \leq m(1), \\
-1, & \text{if } \hat{\alpha}_{\text{LS}} \leq m(-1).
\end{cases}
$$

(5)

Given the observed value of the OLS estimator, say $\hat{\alpha}_{\text{LS}}$, the estimator can be expressed for $m(-1) < \hat{\alpha}_{\text{LS}} \leq m(1)$ as

$$
\hat{\alpha}_{\text{Med}} = m^{-1}(\hat{\alpha}_{\text{LS}}) = \arg\min_{\alpha} |\text{Med}(\hat{\alpha}_{\text{LS}} | \alpha, X) - \hat{\alpha}_{\text{LS}}|.
$$

(6)

Equivalently, with $F_{\hat{\alpha}_{\text{LS}}}$ denoting the cumulative distribution function (cdf) of $\hat{\alpha}_{\text{LS}}$,

$$
m^{-1}(\hat{\alpha}_{\text{LS}}) = \arg\min_{\alpha} |F_{\hat{\alpha}_{\text{LS}}} (\hat{\alpha}_{\text{LS}} | \alpha, X) - 0.5|,
$$

(7)

which is more suitable for computation than (6).

To get an idea of the magnitude of the correction induced by $\hat{\alpha}_{\text{Med}}$, the top panel of Figure 1 plots values of $\hat{\alpha}_{\text{LS}}$ on the ordinate (y-axis) versus the corresponding quantity which should be added to $\hat{\alpha}_{\text{LS}}$ to arrive at $\hat{\alpha}_{\text{Med}}$ on the x-axis (note that the requirement that the median function be monotonic does not imply that the correction, as a function of the observed OLS estimator, is monotonic). For example, with $T = 10$ and an $X$ matrix consisting of intercept and trend, if $\hat{\alpha}_{\text{LS}} = 0.2$, then $\hat{\alpha}_{\text{Med}} \approx 0.68$. As expected, the amount of correction decreases as the sample size increases. One also sees that, particularly for smaller sample sizes, the amount of correction vastly increases when the model changes from intercept to intercept and trend.

Evaluation of (7) is possible and straightforward if $F_{\hat{\alpha}_{\text{LS}}}$ is computable. From expression (13) in the appendix, this involves the cdf of a ratio of quadratic forms in normal variables, which can be evaluated by numerically inverting an associated characteristic function, as detailed by Imhof (1961) in this context. As Andrews (1993) noted, such a computation is prohibitively slow, with simulation being a viable alternative. Because the three special regressor cases (no intercept, intercept, and intercept and time trend) arise frequently in applications, Andrews (1993) tabulated the necessary quantiles, to three significant digits, for a grid of ten (unequally spaced) sample sizes between 40 and 200 and 20 (unequally spaced) $\alpha$ values between $-0.999$ and 1.0. While two-dimensional interpolation of the tabulated values can easily be automated in a computer, it will still result in only about two digit accuracy. Moreover, for sample sizes outside the range [40, 200] or—more likely—a different set of regressors, the tables are not applicable.

To address these shortcomings, use can be made of the so-called saddlepoint approximation, hereafter SPA, for evaluating $F_{\hat{\alpha}_{\text{LS}}}$. The saddlepoint method can be viewed as an accurate approximation to the inversion of the characteristic function, but without the need for numerical
Figure 1: Adjustment to $\hat{\alpha}_{\text{LS}}$ corresponding to $\hat{\alpha}_{\text{Med}}$ (top), $\hat{\alpha}_{\text{Mean}}$ (middle) and $\hat{\alpha}_{\text{Mode}}$ (bottom), shown for the two X matrices constant (solid lines) and constant/trend (dashed lines). The four sample sizes shown are $T = 10$, $T = 25$, $T = 50$ and $T = 100$, moving from right to left. See also the text at the end of Section 2 for explanation.
integration, which gives rise to its enormous speed advantage (see Barndorff–Nielson and Cox, 1989, Jensen, 1995, and Goutis and Casella, 1999, for a general overview, and Lieberman, 1994, for the SPA in the context we use herein). It does, however, require the existence of the moment generating function, which is not true for general random variables, but is available in the setting considered in this paper. An understanding of the SPA is not necessary to use our proposed methods; the appendix contains the required formulae for the computations and relevant references to the literature. Programs (in Matlab) are also available from the authors to perform all the calculations.

As mentioned, the SPA is not exact. It yields two to three digit accuracy for sample sizes between 10 and 30, and from three to four digit accuracy for samples between 50 and 80 (and is asymptotically exact), and so matches (or usually exceeds) that obtained from using linear interpolation from the tables in Andrews (1993). Depending on the method used for numerical integration and the specified tolerance on the error in the Imhof (1961) routine, the SPA is between 10 and 200 times faster, and is easier to program, requiring only a univariate root search and evaluation of the standard normal cdf, both of which are implemented in virtually all statistical computing packages.

4 Mean–adjusted Estimation

For inference on \( \alpha \) via the statistic \( \hat{\alpha}_{LS} \), it appears infeasible to construct a mean–unbiased estimator, but a procedure which comes very close (and turns out to exhibit other good small–sample properties) has been proposed by Tanizaki (2000), and also by MacKinnon and Smith (1998), in a more general context. It amounts to interpreting \( m(\cdot) \) as the analogously defined mean function in (5), i.e., let \( m(\alpha) = \mathbb{E}[\hat{\alpha}_{LS} | \alpha, X] \). Like the median function, numerical results suggest that it is strictly increasing for \( -1 < \alpha < 1 \), so that its inverse exists. In particular, for \( m(-1) < \hat{\alpha}_{LS}^O \leq m(1) \),

\[
\hat{\alpha}_{\text{Mean}} = m^{-1}(\hat{\alpha}_{LS}) = \arg\min_\alpha \left| \mathbb{E}[\hat{\alpha}_{LS} | \alpha, X] - \hat{\alpha}_{LS}^O \right|,
\]

which we refer to as the mean–adjusted estimator. It is not exactly mean–unbiased because of the truncation at \(-1\) and \(1\) and because of the nonlinearity of the mean function, i.e., \( \mathbb{E}[m^{-1}(\hat{\alpha}_{LS})] \neq m^{-1}(\mathbb{E}[\hat{\alpha}_{LS}]) = \alpha \).

Tanizaki used simulation to obtain the mean function \( \mathbb{E}[\hat{\alpha}_{LS}] \) in (8). Due to sampling variation, the inversion of the mean function obtained in that way is prone to instability, rendering a faster and more reliable method for its evaluation of desirable. As such, we suggest to use the expressions for the first and second moments of a ratio of central quadratic forms in normal variables as given in Sawa (1978). Specifically, with \( \mathbf{A} \) and \( \mathbf{B} \) as defined in (13) in the appendix,
let $P \Lambda P'$ be the spectral decomposition of $B$ and set $C = P'AP$. Then

$$E[\hat{\alpha}_{LS}] = \int_0^\infty \sum_{j=1}^{T+1} \frac{c_j}{(1 + 2\lambda_j t)^{3/2}} \prod_{k \neq j} (1 + 2\lambda_k t)^{1/2} dt,$$

(9)

where $c_j$ and $\lambda_j$ denote the $j^{th}$ diagonal element of $C$ and $\Lambda$, respectively. The indefinite integral in (9) can be evaluated directly, by using the fact that most of the integrand mass is near zero and that the integrand dies off rapidly, or by transforming the range of $t$ to lie in an open interval of finite length (e.g., via the substitution $u = 1/(1 + t)$). The former approach was found to be faster and numerically more reliable; see also Paolella (2003).

The middle panel of Figure 1 is similar to the top panel, but shows the correction appropriate for $\hat{\alpha}_{Mean}$. While certainly different, it differs significantly from the top panel only for values of $\hat{\alpha}_{LS}$ less than $-0.4$.

5 Mode–adjusted Estimation

Use of bias adjustment methods based on the mean and median (as measures of central tendency) leads naturally to consideration of the third such measure: the mode. Following (6) and (8), it is natural to define the mode–adjusted estimator as

$$\hat{\alpha}_{Mode} = m^{-1} (\hat{\alpha}_{LS}) = \arg\min_\alpha |\text{Mode} (\hat{\alpha}_{LS} | \alpha, X) - \hat{\alpha}_{O}^{\alpha}|,$$

(10)

where $m(\cdot)$ is now interpreted in (5) as the mode function. To the best of the authors’ knowledge, such an estimator has not been previously proposed. In comparison to $\hat{\alpha}_{Med}$ and $\hat{\alpha}_{Mean}$, which are well–defined and unique for continuous distributions with finite first moment, use of $\hat{\alpha}_{Mode}$ only makes sense if the relevant distribution is unimodal. Indeed, inspection shows that, for sample sizes greater than five, the probability density function (pdf) of $\hat{\alpha}_{LS}$ is unimodal and, paralleling the requirements of the median and mean, the mode function of $\hat{\alpha}_{LS}$ is strictly increasing for $|\alpha| < 1$, thus guaranteeing that $\hat{\alpha}_{Mode}$ is uniquely defined.

Let $f_{\hat{\alpha}_{LS}}(x; \alpha)$ denote the pdf of $\hat{\alpha}_{LS}$ at $x$ when the true parameter is $\alpha$ (and suppress the dependency on the $X$–matrix). From the definition of the mode, it follows that (10) is equivalent to choosing $\hat{\alpha}$ such that the density $f_{\hat{\alpha}_{LS}}(x; \hat{\alpha})$ attains its maximum at the observed value of $\hat{\alpha}_{LS}$. That is, we can write

$$\hat{\alpha}_{LS}^O = \arg\max_x f_{\hat{\alpha}_{LS}}(x; \hat{\alpha}_{Mode}),$$

(11)

i.e., $\hat{\alpha}_{Mode}$ is the (unique) value of $\alpha$ such that the observed value is a mode of $f_{\hat{\alpha}_{LS}}(x; \alpha)$.

Under the stated assumptions of unimodality and monotonicity of the mode of $\hat{\alpha}_{LS}$ as a
function of $\alpha$ (for $|\alpha| < 1$), $\hat{\alpha}_{\text{Mode}}$ is the unique solution to the implicit equation

$$\frac{\partial f_{\hat{\alpha}_{\text{LS}}} (x; \hat{\alpha}_{\text{Mode}})}{\partial x} \bigg|_{x = \hat{\alpha}_{\text{LS}}} = 0,$$

(12)

which can be solved by a univariate root search in much the same fashion as is required for $\hat{\alpha}_{\text{Med}}$ and $\hat{\alpha}_{\text{Mean}}$.

In this context, computation of $\hat{\alpha}_{\text{Mode}}$ is not trivial, as closed–form expressions for $\text{Mode}(\hat{\alpha}_{\text{LS}})$ or $f_{\alpha_{\text{LS}}}(x; \alpha)$ do not exist. Regarding the latter, because the characteristic function of a ratio of quadratic forms is numerically intractable, standard inversion formulae for evaluating the pdf cannot be applied, as is possible for the cdf. We circumvent this problem by again using a saddlepoint approximation: A first and second order SPA to the pdf of a ratio of quadratic forms in zero–mean normal random variables has been constructed by Lieberman (1994). The appendix contains all the relevant expressions.

The one potential drawback of using the SPA for the pdf is that neither the first nor second order expression integrates precisely to one, although both are usually very close (the latter even more so), but the exact integrating constant depends on the true value of $\alpha$ and on the $X$ matrix, and so cannot be determined without numerical integration for each case. Fortunately, for computing $\hat{\alpha}_{\text{Mode}}$, this issue is irrelevant, because the position of the maximum of the density will not change with normalization. Furthermore, the SPA to the pdf is continuous over the (interior of the) entire support (unlike the cdf, which requires some finessing near the mean; see the appendix).

Thus, computation of $\hat{\alpha}_{\text{Mode}}$ can be fully operationalized in this context in a numerically fast and accurate fashion. The only possible remaining caveat to its effective use is the approximate nature of the density via the SPA. To check this, we used the numerical second derivative of the exact cdf of $\hat{\alpha}_{\text{LS}}$, which can be made numerically reliable enough for approximating the mode, but is extremely time consuming compared to use of the SPA. Using the very small sample size of $T = 25$, we found that differences in $\hat{\alpha}_{\text{Mode}}$ based on the SPA and use of the exact cdf occurred only in the third to fourth decimal place, thus confirming that use of the SPA in this context will not jeopardize the accuracy of the method by any appreciable amount. (As $T$ increases, so does the accuracy of the SPA, because the distribution of $\hat{\alpha}_{\text{LS}}$ approaches the normal, for which the SPA is exact.)

The bottom panel of Figure 1 shows the correction appropriate for $\hat{\alpha}_{\text{Mode}}$. Notice how it differs considerably from the other two, implying that its small–sample properties should also differ markedly from those of $\hat{\alpha}_{\text{Med}}$ and $\hat{\alpha}_{\text{Mean}}$. This is indeed the case, and is detailed next.
6 Small Sample Properties of the Point Estimators

6.1 Computation

Having operationalized all three bias–corrected estimators in a fast and accurate fashion, it becomes feasible to conduct a simulation study in order to assess their properties, along with the OLS estimator, and the exact maximum likelihood estimator $\hat{\alpha}_{ML}$. For even greater time savings, the specific simulation scheme used capitalizes on the fact that the three bias–corrected estimators are one–to–one transformations of the least squares estimator; the details can be found in Broda et al. (2004). The corresponding Matlab programs are available from the authors.

The median, mean, and mode–adjusted estimators all have to be truncated above at unity because their existence crucially depends on the invertibility of the median, mean, and mode functions. In order for this condition to hold, the respective functions have to be strictly increasing, which was found to be the case when $|\alpha| \leq 1$. To improve comparability, we chose to restrict $\hat{\alpha}_{LS}$ and $\hat{\alpha}_{ML}$ as well, which anyway mirrors what would be done in practice when working with economic data for which an explosive process ($\alpha > 1$) is untenable. In doing so, observe that, for $\alpha$ values close to or precisely unity, estimators which adjust by larger amounts will be favored.

6.2 Results for a Typical Model

We now discuss in some detail the results for the model with constant and time trend, and $T = 25$ observations (the small sample size being used to help illustrate the differences in the methods; see below for larger $T$). Figure 2 plots the mean bias, median bias and MSE of the various estimators as a function of $\alpha$, computed at $\alpha = -1, -0.8, 0, 0.1, 0.2, \ldots, 0.9, 0.91, 0.92,\ldots, 1$. Regarding mean bias, defined as $E[\hat{\alpha} - \alpha]$, $\hat{\alpha}_{\text{Mean}}$ is indeed the least biased for all values of $\alpha$, but still deviates from zero considerably as $\alpha$ approaches unity. Not surprisingly, for virtually the entire parameter space of interest ($\alpha > -0.5$), $\hat{\alpha}_{\text{LS}}$ is the most biased, drastically so as $\alpha$ approaches unity. Regarding median bias, we confirm that $\hat{\alpha}_{\text{Med}}$ is indeed unbiased while the bias of the other adjusted estimators is not particularly large. (The very small spike in the median bias of $\hat{\alpha}_{\text{Med}}$ near 0.96 is indeed due to use of the SPA for calculation of the relevant cdf. The spike disappears completely for sample sizes $T \geq 35$.)

However, for any estimator to be a useful inferential tool, we require not only that its distribution be centered at the true parameter, but also maximally concentrated around it, as measured naturally by its absolute moments about that value,

$$E \left[ |\hat{\alpha} - \alpha|^d \right], \quad d > 0.$$  

As was stressed above, an estimator with less bias does not necessarily perform better with regard to such a measure, because the bias correction procedure itself may increase dispersion. Taking
Figure 2: Mean bias (top), median bias (middle) and MSE (bottom) of the estimators $\hat{\alpha}_{LS}$ (thin solid line), $\hat{\alpha}_{ML}$ (dotted), $\hat{\alpha}_{Mean}$ (dashed), $\hat{\alpha}_{Med}$ (dash-dot), and $\hat{\alpha}_{Mode}$ (thick solid line) based on a model with constant and time trend, with $T = 25$ observations and normal innovations. Bottom graph is truncated; the MSE of $\hat{\alpha}_{LS}$ increases to 0.16 at $\alpha = 1$. 
For most of the negative \( \alpha \) region, the exact MLE performs best, while for \(-0.1 \leq \alpha \leq 0.7\), \( \hat{\alpha}_{\text{Mode}} \) exhibits the smallest MSE. For \( \alpha > 0.7 \), \( \hat{\alpha}_{\text{Mean}} \) is the best; this result is of particular interest because of the predominance in economic data sets of values of \( \alpha \) near unity. It should be noted, however, that \( \hat{\alpha}_{\text{Mean}} \) exhibits the highest MSE of all the estimators (including \( \hat{\alpha}_{\text{LS}} \)) for \(-0.6 \leq \alpha \leq 0.4\), while for \( 0.4 \leq \alpha \leq 0.7\), \( \hat{\alpha}_{\text{Mean}} \) has the highest MSE among all the bias–corrected estimators. \( \hat{\alpha}_{\text{Med}} \) never achieves the lowest MSE, except at a point near \( \alpha = 0.7 \), where the MSE of all three bias–corrected estimators cross.

### 6.3 Results for other Parameterizations

It must be kept in mind that the previous discussion pertains only to the specific \( X \) matrix and sample size under consideration. Rather conveniently however, it turns out that the results are qualitatively extremely similar for different sample sizes, \( X \) matrices, and distributional assumptions, so that general conclusions can be drawn. This performance is now examined in some more detail.

First consider changing \( X \) to just a column of ones, denoted \( X = 1 \). Figure 3 shows the results for the MSE (the bias results were very similar to those shown in Figure 2 and are omitted). The general shape of the MSE as a function of \( \alpha \) is more hump–shaped, but the ranges of \( \alpha \) for which a particular estimator is preferred are virtually the same. Noticeable is that the MSE of the bias–corrected estimators using \( X = 1 \) does not increase as much as \( \alpha \) approaches unity. Also, \( \hat{\alpha}_{\text{Mean}} \) now exhibits the highest MSE over an even larger range of \( \alpha \), somewhat more than half the parameter space, and \( \hat{\alpha}_{\text{Med}} \) is the second worst for most of the same region.

We now resume use of the constant and trend model with \( T = 25 \) observations, but consider changing the distributional assumption from normal to Cauchy, which possesses tails much fatter than usually arises in empirical applications in econometrics and serves as a special case of both the Student’s \( t \) and symmetric stable Paretian distribution.

It should be kept in mind that the bias–corrected estimators are all based on the normal assumption used in the calculation of the distribution of \( \hat{\alpha}_{\text{LS}} \) in (13). The distribution of quadratic forms in variables other than normal is virtually intractable, though some results on their moments are available; see Roberts (1995), Ullah, Srivastava and Roy (1995) and the references therein. In the case of Cauchy innovations, the asymptotic distribution of the OLS estimator involves the ratio of two independent stable Paretian random variables (Davis and Resnick, 1986). However,
because we truncate the distributions of the estimators at $-1$ and $1$, the mean bias and MSE are still meaningful statistics.

The results are shown in Figure 4. While there are certain differences, the overall behavior of the estimators is still similar to the normal case. For example, $\hat{\alpha}_{\text{Mean}}$ is still approximately unbiased over most of the parameter space, exhibiting an increase in bias as $|\alpha|$ approaches one, as in the normal case. Estimator $\hat{\alpha}_{\text{Med}}$ is no longer median unbiased, but is approximately so for $\alpha < 0.8$. Interestingly, $\hat{\alpha}_{\text{Mode}}$ is also approximately median–unbiased.

The differences in MSE among the bias–corrected estimators are somewhat less pronounced, although, qualitatively speaking, the envelope of minimum MSE is virtually the same, i.e., $\hat{\alpha}_{\text{LS}}$ is recommended over most of the negative $\alpha$ range, $\hat{\alpha}_{\text{Mode}}$ for $-0.1 \leq \alpha \leq 0.7$ and $\hat{\alpha}_{\text{Mean}}$ for $\alpha > 0.7$.

A similar analysis was conducted using other distributional assumptions including Laplace, Student’s $t$ and asymmetric stable Pareto (the latter two with tail indexes such that the mean exists). We chose these distributions (instead of other candidates previously used in similar comparison exercises, such a chi-square or uniform) because their use has become commonplace for capturing the often–observed non–normality of economic and financial data (see, for example, McDonald, 1997; Adler, Feldman and Taqqu, 1998; and Kotz, Podgorski and Kozubowski, 2001, for a vast array of applications). The bias and MSE results were barely distinguishable from those based on a Gaussian assumption, even when using extremely leptokurtic and asymmetric innovations. It appears that, for small sample sizes, the choice of $\mathbf{X}$ has more of an impact than does—even considerable—deviation from normality, in terms of both fatter tails and/or asymmetry.

Figure 3: Same as the bottom panel in Figure 2 but having used only a constant (and no trend) in the regressor matrix.
To further investigate the robustness of our findings, various other \( X \)-matrix specifications were tried, including (i) use of boolean (dummy) vectors, as would be used, for example, when working with models with outliers or structural breaks, and (ii) the matrix \( E_k \), specified as the first \( k \) eigenvectors of the first order difference matrix, for various values of \( k \). Matrix \( E_k \) is given by \( x_{it} = \cos[(2t - 1)\pi(i - 1)/(2T)] \) (see Durbin and Watson, 1971), and is a useful benchmark because these vectors tend to mimic the behavior of economic time series with seasonal and cyclical–type behavior (Dubbelman, Louter and Abrahamse, 1978; King, 1985, p. 32). For all \( X \)-matrices considered, the ranges for which the respective estimators perform best were virtually identical.

Finally, to see the effect of sample size, Figure 5 shows the MSE results for the constant–trend model with normal innovations, but now using \( T = 75 \) observations. As expected, the MSE decreases for all estimators. The \( \alpha \)-ranges and estimators corresponding to the minimum MSE envelope are again virtually the same, but now the difference in MSE of the bias–corrected estimators is far less pronounced. This is expected, because, for \( |\alpha| < 1 \), \( \sqrt{T}(\hat{\alpha}_{LS} - \alpha) \overset{\text{asy}}{\sim} N(0,1 - \alpha^2) \), for which the mean, median and mode coincide. The shape of the MSE curve in Figure 5 is also much closer to \( (1 - \alpha^2)/T \) than those corresponding to \( T = 25 \). Also, because the asymptotic distribution is less accurate for a given sample size as \( \alpha \) approaches one, the discrepancy in MSE shown in Figure 5 increases as \( \alpha \) approaches one (with \( \hat{\alpha}_{\text{Mean}} \) exhibiting the lowest MSE).

7 Conclusions

The median–unbiased point estimator suggested by Andrews (1993) for the first order autoregressive coefficient is statistically well–motivated and also computationally feasible, owing to the tabulated values provided by Andrews. Nevertheless, its implementation based on these tables was restricted to models with either no intercept, intercept, or intercept and trend, while many models will require different exogenous regressors, including dummies to pick up structural breaks or outliers, thus requiring time–consuming custom calculations. This paper uses a saddlepoint approximation to the required distribution function, rendering such custom calculations feasible as a routine task. The mean–adjusted estimator of Tanizaki (2000) suffers from a similar computational burden, owing to the use of simulation in the construction of the mean function. We have shown how this time–consuming process can be replaced with a fast, exact calculation.

The availability and high accuracy of the saddlepoint approximation facilitates other computationally intensive estimation methods. In particular, we propose a new estimator which uses the mode as a measure of central tendency. This estimator takes under a second to compute (when using the density saddlepoint approximation), and does not appear to have been entertained in any statistical modeling context previous to this study.
Figure 4: Same as Figure 2 but having used Cauchy innovations
With all the estimators computable in a fast and reliable fashion, a simulation study comparing their small–sample properties becomes feasible. With respect to MSE, we demonstrate that \( \hat{\alpha}_{\text{Mode}} \) is superior for the large part of the parameter space, \(-0.1 < \alpha < 0.7\), while \( \hat{\alpha}_{\text{Mean}} \) is shown to exhibit lower MSE for the important region \( 0.7 < \alpha < 1 \).

Perhaps the most important observation regarding these results on MSE is that they appear to hold almost exactly irrespective of the choice of exogenous regressors, sample size, and innovation distribution assumption. Concretely, this means that an estimator with relatively lowest MSE (of the estimators entertained herein) can be selected, assuming correct prior opinion of whether \( \alpha < -0.1 \), \(-0.1 < \alpha < 0.7\) or \( 0.7 < \alpha < 1 \), the latter being a common choice.

Software for Matlab is available from the authors (i) to compute the estimators discussed herein for a given data set, and (ii) to determine the properties of the point estimators over a grid of values of the autoregressive parameter for any exogenous regressor matrix and choice of innovation assumption (included are normal, Student’s \( t \), stable Pareto and Laplace, though others are easily incorporated). This can be used to help decide on a suitable point estimator to report based on the observed data, although, as discussed above, our simulations show that the optimal choice of estimator is virtually invariant to the regressor matrix, sample size, and innovations assumption.

Given the ease of computation, one could entertain a two step procedure when faced with no prior information: First estimate \( \alpha \) with, say, \( \hat{\alpha}_{\text{Mode}} \), which exhibits good overall MSE performance, and, based on it, the estimator with lowest MSE is selected to deliver the final estimate. The small sample performance of such a procedure might be worth investigating, although, realistically, most researchers will have, to some extent, a prior on \( \alpha \).

The estimation techniques developed herein could be applied to models for panel data. For example, Phillips and Sul (2003) make use of the median unbiased estimator in the context of dynamic panel models. Work in progress includes extending the methodology to the AR(\( p \)) and ARMA(\( p, q \)) case which complements the method for approximate median unbiasedness in the AR(\( p \)) model developed by Andrews and Chen (1994).

**Appendix: Saddlepoint Approximation to the cdf and pdf of (4)**

\( \hat{\alpha}_{\text{LS}} \) in (4) is independent of \( \beta \) and \( \sigma^2 \) (and of \( Y_0 \) if \( \alpha = 1 \)) for any exogenous regressor matrix \( X \), so without loss of generality, we can assume \( \beta = 0 \) and \( \sigma^2 = 1 \) in the following. Defining the selection matrices \( D_T = [0 \mid I_T] \) and \( D_{T-1} = [I_T \mid 0] \), we have \( MY_T = MD_T Y^\ell \) and \( MY_{T-1} = MD_{T-1} Y^\ell \), where \( Y^\ell = [Y^\ell_0, (Y^\ell_T)^\prime] \). Substituting this into (4) yields

\[
\hat{\alpha}_{\text{LS}} = \frac{(Y^\ell_T)^\prime D_{T-1}' MD_T' Y^\ell}{(Y^\ell_T)^\prime D_{T-1}' MD_{T-1}' Y^\ell} = \frac{U'R'D_{T-1}' MD_T' RU}{U'R'D_{T-1}' MD_{T-1}' RU} = \frac{U'AU}{U'BU},
\]

(13)
cumulant generating function of $S$ chi-squared with one degree of freedom and the $\lambda$ given by

This needs to be numerically solved. Daniels (1987) derived the next term in the expansion (16), normal distribution, respectively, and $\hat{s}$ where $\hat{s}$ is the (unique) saddlepoint which satisfies

The general result of Lugannani and Rice (1980) can be directly applied to $S$ to yield the saddlepoint approximation to its cdf; it is

where $\hat{\alpha}$ is the so-defined weighted sum, $b = (1 - \alpha^2)^{-1/2}$ if $\alpha \in (-1, 1)$ and zero if $\alpha = 1$. If there are no exogenous regressors, set $M = I$ and all conditioning on $X$ is replaced by conditioning on $T$. Hence, for given $\alpha$, the cdf of $\hat{\alpha}$ can be expressed as

\[
\Pr (\hat{\alpha}_{LS} \leq c) = \Pr \left( U' \left( A/2 + A'/2 - cB \right) U \leq 0 \right) = \Pr \left( U' W U \leq 0 \right),
\]

with symmetric matrix $W = W (\alpha, c)$ so defined. From the principle axis theorem,

where $S$ is the so-defined weighted sum, $w = \text{rank}(W)$, each $\chi_i^2(1, 0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the $\lambda_i$ are the eigenvalues of $W$. Let $K = K_S$ be the cumulant generating function of $S$, given by $K_S(s) = \frac{1}{2} \sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.

The general result of Lugannani and Rice (1980) can be directly applied to $S$ to yield the saddlepoint approximation to its cdf; it is

\[
\tilde{F}_S(x) = \Phi (\hat{w}) + \phi (\hat{w}) \left\{ \frac{1}{\hat{w}} - \frac{1}{\hat{\alpha}} \right\}, \quad x \neq E[S],
\]

where $\hat{\alpha}_{LS} \leq c = \Pr \left( U' \left( A/2 + A'/2 - cB \right) U \leq 0 \right) = \Pr \left( U' W U \leq 0 \right),
\]

with symmetric matrix $W = W (\alpha, c)$ so defined. From the principle axis theorem,

where $S$ is the so-defined weighted sum, $w = \text{rank}(W)$, each $\chi_i^2(1, 0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the $\lambda_i$ are the eigenvalues of $W$. Let $K = K_S$ be the cumulant generating function of $S$, given by $K_S(s) = \frac{1}{2} \sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.

The general result of Lugannani and Rice (1980) can be directly applied to $S$ to yield the saddlepoint approximation to its cdf; it is

\[
\tilde{F}_S(x) = \Phi (\hat{w}) + \phi (\hat{w}) \left\{ \frac{1}{\hat{w}} - \frac{1}{\hat{\alpha}} \right\}, \quad x \neq E[S],
\]

where $\hat{\alpha}_{LS} \leq c = \Pr \left( U' \left( A/2 + A'/2 - cB \right) U \leq 0 \right) = \Pr \left( U' W U \leq 0 \right),
\]

with symmetric matrix $W = W (\alpha, c)$ so defined. From the principle axis theorem,

where $S$ is the so-defined weighted sum, $w = \text{rank}(W)$, each $\chi_i^2(1, 0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the $\lambda_i$ are the eigenvalues of $W$. Let $K = K_S$ be the cumulant generating function of $S$, given by $K_S(s) = \frac{1}{2} \sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.

The general result of Lugannani and Rice (1980) can be directly applied to $S$ to yield the saddlepoint approximation to its cdf; it is

\[
\tilde{F}_S(x) = \Phi (\hat{w}) + \phi (\hat{w}) \left\{ \frac{1}{\hat{w}} - \frac{1}{\hat{\alpha}} \right\}, \quad x \neq E[S],
\]

where $\hat{\alpha}_{LS} \leq c = \Pr \left( U' \left( A/2 + A'/2 - cB \right) U \leq 0 \right) = \Pr \left( U' W U \leq 0 \right),
\]

with symmetric matrix $W = W (\alpha, c)$ so defined. From the principle axis theorem,

where $S$ is the so-defined weighted sum, $w = \text{rank}(W)$, each $\chi_i^2(1, 0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the $\lambda_i$ are the eigenvalues of $W$. Let $K = K_S$ be the cumulant generating function of $S$, given by $K_S(s) = \frac{1}{2} \sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.

The general result of Lugannani and Rice (1980) can be directly applied to $S$ to yield the saddlepoint approximation to its cdf; it is

\[
\tilde{F}_S(x) = \Phi (\hat{w}) + \phi (\hat{w}) \left\{ \frac{1}{\hat{w}} - \frac{1}{\hat{\alpha}} \right\}, \quad x \neq E[S],
\]

where $\hat{\alpha}_{LS} \leq c = \Pr \left( U' \left( A/2 + A'/2 - cB \right) U \leq 0 \right) = \Pr \left( U' W U \leq 0 \right),
\]

with symmetric matrix $W = W (\alpha, c)$ so defined. From the principle axis theorem,

where $S$ is the so-defined weighted sum, $w = \text{rank}(W)$, each $\chi_i^2(1, 0)$, $i = 1, \ldots, w$, are iid central chi-squared with one degree of freedom and the $\lambda_i$ are the eigenvalues of $W$. Let $K = K_S$ be the cumulant generating function of $S$, given by $K_S(s) = \frac{1}{2} \sum_{i=1}^{w} \ln v_i$, where $v_i = 1/(1 - 2s\lambda_i)$.
of values in a small neighborhood of $x$.

A second order SPA to the pdf of $\hat{\alpha}_{LS}$ is developed in Lieberman (1994). With $A, B, W(\alpha, c)$ and $v_i, i = 1, \ldots, w$, as previously defined, the first-order approximation is given by

$$f_{\hat{\alpha}_{LS}}(c) = \frac{\text{tr}[ (I - 2\hat{s}W)^{-1}B] \exp \{ \frac{1}{2} \sum_{i=1}^{w} \ln v_i \}}{\sqrt{4\pi \sum_{i=1}^{w} (\lambda_i v_i)^{-2}}},$$

(18)

where $\hat{s}$ denotes the same unique saddlepoint as is used in approximating the cdf.

Expression (18) is the leading term in an asymptotic expansion; the second order approximation is given by

$$f_{\hat{\alpha}_{LS}}(c) = \tilde{f}_{\hat{\alpha}_{LS}}(c) \left( 1 - \frac{2 \text{tr} (K^2 L)}{\text{tr}(L) \text{tr}(K^2)} + \frac{3 \text{tr} (K^4)}{2(\text{tr} K^2)^2} + \frac{2 \text{tr} (KL) \text{tr}(K^3)}{\text{tr}(L)(\text{tr} K^2)^2} - \frac{5(\text{tr} K^3)^2}{3(\text{tr} K^2)^3} \right),$$

(19)

where $K = \hat{H}^{-1}W$, $L = \hat{H}^{-1}B$ and $\hat{H}^{-1} = \hat{H}^{-1}(\hat{s}) = I - 2\hat{s}W$. Calculations reveal that the expression in Lieberman’s Equation (5) contains a misprint; it is correct as given here.

Acknowledgements

Part of the research of M. Paolella has been carried out within the National Center of Competence in Research “Financial Valuation and Risk Management” (NCCR FINRISK), which is a research program supported by the Swiss National Science Foundation.

The authors are grateful to Donald Andrews, Jörg Breitung, Uwe Hassler, the participants of the Workshop on Risk Analysis in Finance and Insurance in June 2004, the Ausschuß für Ökonometrie (Verein für Socialpolitik) in March 2005, two anonymous referees, and the co-editor, Erricos Kontoghiorghes, for valuable comments and suggestions.
References


Figure 5: Same as the bottom panel of Figure 2 but using $T = 75$