Measurement of the Quality of Autoregressive Approximation, with Econometric Applications

John W. Galbraith and Victoria Zinde-Walsh
Department of Economics
McGill University
855 Sherbrooke St. West
Montreal, Quebec H3A 2T7 CANADA

Abstract

A distance measure in the appropriate space of stochastic processes can be used to measure the quality of approximation when one process is taken as a model of another, either deliberately or by mis-specification. We examine the problem of approximating an ARMA process by a model from the AR(p) class, emphasizing a distance measure based on the Hilbert metric. This measure can be used to calculate distances between particular processes, and the minimum distance to a class of processes such as the AR(p) class. We show that this measure provides a good *a priori* indication of the impact of substitution of an approximate process for the true process. We also provide comparison with the Kullback-Leibler-Jeffreys information metric, and applications to choice of order in selecting an approximating AR model on a finite sample, testing of dynamic specification, forecast performance of approximate models, and evaluation of information criteria for selection of approximating models.

Key words: autoregressive approximation, ARMA process, Hilbert distance, Kullback-Leibler-Jeffreys distance

JEL Classification nos.: C22, C52

I. Introduction

There are many circumstances in which one stochastic process is taken as a model of another, either inadvertently through mis-specification, or deliberately as an approximation. In the present paper we are concerned with cases in which an autoregressive–moving average (ARMA) or moving-average process is, explicitly or implicitly, approximated by a pure autoregressive process. It is well known (see for example [1]) that an ARMA process with all latent roots of the moving-average polynomial inside the unit circle can be approximated arbitrarily well by an autoregressive process of order ℓ, as $\ell \to \infty$. The technique has been used for the estimation of moving-average or ARMA models by, among others, [2], [3], [4], [5], [6], [7] and [8]. References [9] and [10] address the estimation of the spectral density through autoregression; [11] uses autoregression to approximate an ARMA error process in the residuals of a regression model, and [12] addresses the impact of that approximation on the asymptotic distribution of the ADF statistic.

In problems such as these, the quality of the approximation affects some statistic of interest, and an ideal measure of the quality of the approximation would be monotonically related to the deviations caused by replacing the true process by the approximate one. As an example of the use of such a measure, consider a forecast based on a mis-specified model. If the accuracy of the forecast is monotonically related to some measure of the divergence between the true and mis-specified models, one can make an immediate use of the divergence measure in designing Monte Carlo experiments to evaluate forecast performance for different models and types of mis-specification; the measure allows us to identify cases where the approximation will do relatively well or badly, and to be sure of examining both.

The present study presents an approach to problems of this type. We treat autoregres-

sive approximation and mis-specification in a common framework, implying replacement of the true model with one from another class, and use distance in the space of stationary stochastic processes as a measure of the severity of mis-specification, or quality of approximation. A measure of the distance from a process to a class of processes is defined, and may be minimized to find the closest member of that class.[1] We are able to indicate the order of AR process necessary to approximate particular MA(1) or MA(2) processes well, and are also able to give some general results on the value of the distance between processes as an indicator of the adequacy of an approximation in particular circumstances. For MA(1) processes the magnitude of the root is often mentioned as the factor determining the degree to which autoregressive approximation will be successful; here we are able to give a more general result.

It is important to distinguish these results about the appropriate order of approximating process from the use of sample-dependent criteria such as the Akaike or Schwarz information criteria to choose the order. While the two approaches may to some extent be complementary, the present study offers *a priori* information about the ability of an AR(ℓ) process, for given ℓ, to approximate a particular ARMA. In empirical applications, this information may be combined with information about the process being approximated, and a loss function, to generate a specific choice of order. Distance measures may also be used to evaluate information criteria in particular contexts, as in the example of section IV.D.

We offer several other econometric applications in section IV. The distance measure is defined and described in section II, while section III discusses its use in examining AR approximations.

[1]There are various possible measures of distance or divergence, including the well-known Kullback-Leibler and Hilbert distances; we concentrate here on the latter.

II. Definitions and properties of the distance measures

This section concentrates on the distance measures, particularly the Hilbert distance, which will be used for the problem of autoregressive approximation. For general reviews of information theory and distance measures, see [13] and [14].

We consider a discrete-time stochastic process $\{X_t\}$. The space of zero-mean, finite-variance stochastic processes can be represented as a real Hilbert space H with the scalar product (X, Y) defined by $E(XY)$; the Hilbert norm $\|X\|$ is given by $[E(X^2)]^{1/2}$. The values of the stochastic process $\{X_t\},\ t \in Z$ (where the index set Z is the set of integers), span a subspace $H_x \subset H$ of the Hilbert space, which is itself a separable Hilbert space and thus has a countable basis. The lag operator L is defined such that $LX_t = X_{t-1}$; [15] and [16], for example, describe the relevant definitions and properties of the stationary stochastic processes and the Hilbert spaces used here. For the purpose of examining mis-specification, we restrict ourselves to the space H_x.

II. A. The Hilbert distance

The Hilbert distance is the primary measure that we will use.

Since the space of second-order stationary stochastic processes is a Hilbert space, the *distance* between two processes X and Y is given by the norm of the difference, $d_H(X, Y) = \|X - Y\| = [E(X - Y)^2]^{1/2}$. In [17], this distance is used to examine mis-specification in first-order processes. In a Hilbert space, we can easily define the distance from a process to a *class* of processes (or the distance between classes), obtained by minimizing the distance over all processes in the class: for example, for the distance to the AR(ℓ) class, $d_H(X, AR(\ell)) = \inf_{Y \in AR(\ell)} d_H(X, Y)$.

The distance can also be expressed using the innovations representation of the processes in terms of the stationary uncorrelated process: i.e., the orthogonal basis $\{e_t\}_{-\infty}^{\infty}$.

If $X_t = \sum_{i=0}^{\infty} \beta_i e_{t-i}$ and $Y_t = \sum_{i=0}^{\infty} \xi_i \varepsilon_{t-i}$ with $\varepsilon_t = \sigma e_t$, then

$$d_H(X,Y) \ = \|X - Y\| \ = \left[\sum_{i=0}^{\infty} (\beta_i - \sigma\xi_i)^2\right]^{1/2} \sigma_e, \tag{2.1}$$

where $\|e_t\| = \sigma_e$, the standard error of the $\{e_t\}$. Without loss of generality we will consider $\sigma_e = 1$ below unless otherwise specified.

We will consider processes that can be represented as $X_t = f(L)e_t$, where e_t is a white noise process and $f(L)$ is a rational polynomial, so that $f(L) = Q(L)/P(L)$ where Q and P are polynomials; we will express this as $P(L) \ = I \ - \alpha_1 L \ - \ \cdots \ - \ \alpha_p L^p$, and $Q(L) \ = I \ + \theta_1 L \ + \cdots + \ \theta_q L^q$. An ARMA(p,q) process is described by $P(L)X_t = Q(L)e_t$, and is stationary if and only if the latent (i.e. inverse) roots of the polynomial $Q(L)$ are within the unit circle. If the process is invertible, then the inverse process $\{X_t^-\}$ defined by $Q(L)X_t^- = P(L)e_t$ is stationary. It is normally assumed that $P(L)$ and $Q(L)$ have no common factors. If $P(L) \equiv I$ then $\{X_t\}$ is an MA process; if $Q(L) \equiv I$, it is an AR process.

A stationary, zero-mean ARMA(p,q) process $\{X_t\}$ can be approximated arbitrarily well by an MA(k) process for some k: for an arbitrary bound δ on the approximation error, fix k such that $\sum_{i=k+1}^{\infty} \beta_i^2 < \delta$, and set the parameters θ_i of the approximating MA(k) process $\{Y_t\}$ such that $\theta_i = \beta_i$ for $i = 1, \ldots, k$. It follows that $\|X - Y\| < \delta^{1/2}$. If $\{X_t\}$ is an invertible process, then for sufficiently large k, $\{Y_t\}$ will also be invertible.

Moreover, if $\{X_t\}$ is invertible then it is also possible to express X_t as a convergent weighted sum of past values X_{t-i}, so that we can also find an AR(ℓ) process which approximates $\{X_t\}$ arbitrarily well. Consider an invertible kth-order moving-average lag polynomial represented by $Q_k(L)$, corresponding to $\{Y_t\}$ above. It has an infinite AR representation with autoregressive polynomial $P_k(L) \equiv [Q_k(L)]^{-1}$. If $Q_k(L) = I + \theta_1 L + \cdots + \theta_k L^k$, then $P_k(L) \ = (I \ + \ \theta_1 L \ + \ \cdots \ + \ \theta_k L^k)^{-1} = I \ - \ \theta_1 L \ + \ (\theta_1^2 \ - \ \theta_2 \ ) \ L^2 \ + \ \cdots = \sum_{i=0}^{\infty} \gamma_i L^i$.

Denoting by ν_i the latent (i.e. inverse) roots of $Q_k(L)$, note that $\gamma_i \approx O(\bar{\nu}^i)$, where $\bar{\nu} = \max_{1\le i\le k} |\nu_i|$, and $|\ .\ |$ represents the modulus of the root. Thus $\sum_{\ell+1}^{\infty} \gamma_i^2 \approx O(\bar{\nu}^{2\,\ell+2})$, and for suitable order ℓ of the approximating process, this can be made less than any chosen δ. Denoting by $\{Z_t\}$ the AR(ℓ) process with coefficients $\alpha_i = \gamma_i, i = 1, \ldots, \ell$, we have $\|X-Z\| = \|X-Y+Y-Z\| \le \|X-Y\| + \|Y-Z\| = \left(\sum_{k+1}^{\infty} \beta_i^2\right)^{1/2} + \left(\sum_{\ell+1}^{\infty} \gamma_i^2\right)^{1/2}$. Hence an AR($\ell$) process can be found which is arbitrarily close to $\{X_t\}$ in the Hilbert metric. Also, convergence in the Hilbert metric implies convergence of the Fourier coefficients of the representation in the orthogonal basis of the processes.

As an example, consider an invertible MA(q) process $X_t = e_t + \sum_{j=1}^{q} \theta_j e_{t-j}$ with $\mathrm{var}(e_t) = 1$, which is approximated by the AR(p) process $Z_t = \sum_{j=1}^{p} \alpha_j Z_{t-j} + \varepsilon_t$ with $\mathrm{var}(\varepsilon_t) = \sigma^2$, that minimizes the Hilbert distance. As $p \to \infty$, the Hilbert distance between $\{X_t\}$ and $\{Z_t\}$ approaches zero, $\sigma \to 1$, and the first q coefficients $\{\alpha_j\}$ approach the values $\alpha_1 = \theta_1,\ \alpha_2 = -\theta_1\alpha_1 + \theta_2,\ \alpha_i = -\theta_1\alpha_{i-1} - \theta_2\alpha_{i-2} - \cdots - \theta_{i-1}\alpha_1 + \theta_i$ for $i \le q$, and $\alpha_j = \sum_{i=1}^{q} -\theta_i\alpha_{j-i}$ for $j \ge q+1$. These relations are used for parameter estimation in [7] and [8].

The Hilbert distance between second-order stationary processes in H corresponds to convergence in probability in that class. In fact, since it is defined through the mean square, convergence in this metric implies convergence in probability. On the other hand, convergence in probability to a process in H implies that the processes converge in mean square. Of course, if the processes in H converge in probability to a non-stationary process, they do not converge in this metric. The correspondence to convergence in probability makes the Hilbert metric a valuable measure of "closeness" in the space H, which can be used to evaluate the quality of various approximations. Unlike measures in finite parameter spaces, this measure can be used to compare processes of different types and orders.

II. B. The Kullback-Leibler and Kullback-Leibler-Jeffreys divergence measures

These two divergence measures are based on information functionals; see for example [18] or the review in [14]. For the Shannon entropy functional the Kullback-Leibler (K-L) divergence from a distribution with a density function $\phi_1(y)$ to a distribution with density $\phi_2(y)$ is given by

$$I(\phi_1 : \phi_2) = \int [\log(\phi_1/\phi_2) - 1]\, \phi_1 dy.$$

This measure of divergence is not symmetric; it is sometimes called directional. The Kullback-Leibler-Jeffreys(K-L-J) divergence measure is non-directional (symmetric) and is defined as

$$d_{KLJ}(\phi_1, \phi_2) = \frac{1}{2}\left[I(\phi_1 : \phi_2) + I(\phi_2 : \phi_1)\right].$$

Note that although symmetric, D is not a distance since it does not satisfy the triangle inequality. For Gaussian processes $X_t = f_1(L)e_t$ and $Y_t = f_2(L)e_t$, these divergence measures can be calculated as

$$I(X : Y) = (2\pi)^{-1} \int_0^{2\pi} \left[f_1(e^{iw}) f_1(e^{-iw}) f_2^{-1}(e^{iw}) f_2^{-1}(e^{-iw}) - 1\right] dw,$$

and we can compute $I(Y : X)$ similarly; then

$$d_{KLJ}(X, Y) = \frac{1}{2}\left[I(X : Y) + I(Y : X)\right]. \tag{2.2}$$

The Hilbert distance can be represented through f_1, f_2 as $\|X - Y\|^2 =$

$$(2\pi)^{-1} \int_0^{2\pi} \left[f_1(e^{iw}) f_1(e^{-iw}) + f_2(e^{iw}) f_2(e^{-iw}) - f_1(e^{iw}) f_2(e^{-iw}) - f_2(e^{iw}) f_1(e^{-iw})\right] dw.$$

We can also represent $d_{KLJ}(X, Y)$ via the Hilbert norm. If we define a process $\{Z_t\}$ via $Z = f_1(L)/f_2(L)e = \sum_i \omega_i e_{t-i}$, and define $\overline{Z} = f_2(L)/f_1(L)e = \sum_i \overline{\omega}_i e_{t-i}$, then

$$d_{KLJ}(X, Y) = \frac{1}{2}\left[\|Z\|^2 + \left\|\overline{Z}\right\|^2\right] - 1, \tag{2.3}$$

where $\|Z\|^2 = \sum \omega_i^2$ and $\left\|\overline{Z}\right\|^2 = \sum \overline{\omega}_i^2$. The formula (2.3) can be used to compute the Kullback-Leibler-Jeffreys divergence from one process to another and can be minimized over a particular class to find the minimum divergence from a given process to a class of processes. While our primary focus in this paper is on the use of the Hilbert distance, we will incorporate K-L-J distance measures into several examples below for purposes of comparison.

Before addressing some applications of these concepts, we note that it may be useful to restrict somewhat the class of mis-specified models considered in the applications. We may assume that some characteristics will be shared between the true and mis-specified models; in particular, if we know that some moments exist, we may wish to consider a mis-specified process with the same moments. Indeed, if we were to use moments in the estimation they would come from the same time series data regardless of which model was specified. Since stationary stochastic processes possess at least two moments, here we consider as the approximation the closest process in the approximating class, subject to the restriction that the first two moments are the same as those of the process being approximated. We apply this restriction below in using both Hilbert and K-L-J distances. The K-L-J distance then becomes

$$d_{KLJ}(X,Y) = \frac{1}{2}\left[\sum \omega_i^2(v_2/v_1) + \sum \overline{\omega}_i^2(v_1/v_2)\right] - 1,$$

where v_1, v_2 are the variances of the processes X_t and Y_t defined above. In the case of the Hilbert distance, we normalize one of the sets of squared projection coefficients by the ratio of variances.

III. Evaluation of approximations using distance measures

When we use techniques that approximate one process by a process from another class, we can identify some member or members of the approximating class that are closest to the

original process, by the Hilbert (or other) distance. We will refer to the distance between the original process and an approximating process in a given class as the *approximation distance,* and will be interested in calculating the minimum approximation distance achievable.[2] As discussed in Section II.C, the approximate process is restricted to have the same mean and variance as the original process.

In order to evaluate this minimum (Hilbert) approximation distance, express the original process and a candidate approximating process in terms of the projections onto past innovations. The function describing the distance between them, (2.1), is the sum of squared differences between the coefficients of these innovations representations. Truncating this expression at a large value, the distance may be calculated, and with subsequent iterations the function can be minimized numerically over the parameters of the approximating process. In the calculations below we use a Powell algorithm (see [20]: 299) to minimize the distance function.

Tables 1 and 2 give these examples of the approximation distances from specific invertible MA(1) and MA(2) processes to the closest members of the AR(p) class, $p = 1, 2, 4, 8, 12$; the approximating process is constrained to have the same variance as the original process. Table 2b gives the parameter values and roots of the processes appearing in Table 2a. These distances cannot be guaranteed to be global minima, but appear to be very close to them, at least for distances on the order of 10^{-8} or greater. The tables also report the distances from the original processes to the uncorrelated, or white noise, process having the same variance. For MA(1) processes, the distances are unaffected by

[2] [19] discusses a related concept, the approximation bias arising from the use of a finite-order AR(p) in place of the AR(∞) representation of a process. Parzen introduces a particular penalty function with which to estimate the approximating order, yielding the *criterion of autoregressive transfer function* for order selection.

the sign of the parameter. While for MA(1) processes the distance is a monotonic function of the modulus of the root, note that this is not the case with respect to the largest root of MA(2) processes.

These examples suggest at least two conclusions. First, through most of the MA(1) or MA(2) parameter spaces, the approximation distance can be made quite small with moderate orders of approximating process. For MA(1) processes, order 8 is sufficient in all cases to make the approximation distance less than 1% of the distance of the original process to the uncorrelated process (that is, the approximation has picked up 99% of the original process, by our distance measure). For the MA(2) processes used in these examples, order 12 is sufficient in most cases to meet the same condition, but is not sufficient in cases 1,2,3,4 and 7, where there is one or more root with modulus greater than 0.85 in absolute value. Nonetheless in most cases it is clearly possible to make the approximation distances very small with orders of AR process that are well within the range estimable with typical samples of data.

Second, these results give an *a priori* indication of the appropriate order of approximating AR process. For moving average processes with the largest root near zero, there is little gain in increasing the order, p, beyond fairly small values. For processes with a root near the boundary of the invertibility region, there are still substantial gains in increasing p beyond 12, and the order of AR process necessary to make the approximation distance negligible may be large. This requirement imposes a lower bound on the sample size necessary to provide a good approximation with an autoregressive process.[3] Note, however, that these results do not embody the effect of increased model order on efficiency of parameter

[3] As well, small reductions in approximation distance become more important with increasing sample size, since overall distance from the estimated representation to the true process is itself declining in expectation.

estimation; results bearing on this question are presented in Section IV.A.

[Tables 1, 2a, 2b about here]

The magnitude of approximation distance that is tolerable will depend upon the application. Nonetheless it is worth emphasizing that this information about the order of the approximating process is not sample dependent. It is well known that widely-used sample-based criteria for order selection, such as the Akaike Information Criterion, may systematically suggest over- or under- parameterization; see [21] and the examples in section IV.D below. A criterion such as the distance in the space of population models, by contrast, provides a guide to order selection prior to estimation.

IV. Econometric applications

There are two types of problem which we can distinguish as being of interest in the context of mis-specified or approximate models. In the first type, the statistic is directly related to the mis-specification, and an example is given in section IV.B, where we examine a test for the null of uncorrelated residuals in a model where MA errors are modelled by autoregression. In the second type, a statistic may estimate or test some property not directly related to the mis-specification; the mis-specification is nonetheless relevant because the distribution of the statistic will differ from the distribution that it would have with a correctly specified model. Examples are given in section IV.C, where we consider the forecast error arising when MA processes are forecast using AR models, and in IV.D, where we examine the performance of information criteria in selecting the order of model which is the best approximation to an unknown process of more general form.

In each of these cases, we expect that the more severe the mis-specification, or the poorer the approximation, the more substantial will be the effect on the statistic of interest. Ideally, we would like to have a measure of the extent of mis-specification which has

predictive power in a wide variety of circumstances. In the examples just mentioned, this would allow us to predict which of various MA processes will show the higher mean squared forecast error when forecasting is done via an AR, or which MA process in the errors of a regression model will lead to the largest average test statistic in an autocorrelation test when modelled as AR. In these examples, we show that the Hilbert distance performs well as a such a measure, and in particular, that it is a much better indicator than is the largest of the moduli of MA roots. While the Hilbert distance is the primary focus of our interest, we will also refer for comparison to the Kullback-Leibler-Jeffreys distance in two of the applications.

Before exploring these examples, in which the Hilbert distance measure is used to predict the values of sample-based criteria and thereby evaluated, we apply this distance measure directly to the general problem of choice of AR order, which forms an element of the examples in IV.B, IV.C and IV.D.

IV. A. Choice of AR order

Misspecification or approximation can be thought of as yielding two sources of error: one caused by the mismatch between the mis-specified process (e.g., x) and the true process (e.g. y), and the other resulting from estimation of the mis-specified model (yielding $\hat{x}$ rather than x. Each of these, the approximation error and estimation error, play a role in determining the best approximating process, as the following application illustrates.

Consider the estimation of an AR model of a pure MA process. In choosing the best order for the AR model, there are two offsetting effects: first, as section III showed, the best available approximation within the AR(k) class will be closer to the true process as k increases; second, as k increases the efficiency of parameter estimation will be reduced, leading to a higher mean distance to the true process. We will use the Hilbert distance

to investigate the optimal model order, $k^* = \text{argmin}_{k:\hat{x}\in AR(k)} E(\|y - \hat{x}\|)$, given these two effects.

For a given process y, and an approximating AR(k) model, there is a closest process x within the AR(k) class, and an estimated model $\hat{x}$. As k increases, x becomes a better approximation by the Hilbert distance ($\|y - x\|$ decreases monotonically). Parameter estimation becomes less efficient, however, and the mean distance of the estimated model to the best approximating model, $\|\hat{x} - x\|$, increases. The overall distance between true and estimated processes, $\|y - \hat{x}\|$, will have a minimum at some finite value of k.

Figures 1 to 3 present the results of simulations designed to estimate the relation between $\|y - \hat{x}\|$ and k for several examples of MA processes. There are 10,000 replications on sample sizes of $T = \{200, 1000\}$, and $k = \{1, 2, \ldots, 10\}$. Values on the vertical axis are the average values of $\|y - \hat{x}\|$ for the given MA process, y, across the 10,000 samples.

Note first that the optimal order increases in T, reflecting diminished relative importance of parameter estimation error, at a given k, as T increases. Optimal order also increases, subject to the integer constraint on k, as the distance between the true process and the closest process in the AR(k) class increases (see, again, Tables 1 and 2a). For $\theta = 0.90$, optimal orders are 5 and 9 at $T = 200$ and 1000, for $\theta = 0.50$, optimal orders are 3 and 4, while for $\theta = 0.10$ there is no gain in approximating with an order greater than 1 at either sample size.

These results are purely illustrative. However, we can summarize the results of a larger number of such experiments by estimating a response surface for the optimal order as a function of the parameter of an MA(1) model and sample size, with $\theta = \{0.05, 0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 0.95, 0.99\}$ and $T = \{25, 50, 75, 100, 200, 300, 400, 500, 1000\}$, yielding 81 cases. The response surface (standard errors in brackets)

$$k^* = \text{-2.82} \quad +4.67(1-\theta)^2 \quad \text{-0.23 } T^{1/4} \quad +2.56\ \theta T^{1/4} + \text{u}$$

$$(0.17) \qquad (0.32) \qquad\qquad (0.06) \qquad\quad (0.08)$$

was found to provide a reasonable fit ($R^2 = 0.98$) to the points. For example, using the processes examined in Figures 1–3, we have for each combination (θ, T) the following estimated optimal orders: (0.1,200), $\hat{k}^* = 1.07$; (0.1,1000), $\hat{k}^* = 1.12$; (0.5,200), $\hat{k}^* = 2.30$; (0.5,1000), $\hat{k}^* = 4.26$; (0.9,200), $\hat{k}^* = 5.02$; (0.9,1000), $\hat{k}^* = 8.89$. Each of these is quite close to the actual optimum for the given (θ, T).

IV. B. Dynamic specification

Appropriate specification of dynamics is an important problem in time series regression; see Hendry [22]. for a thorough review of this literature. One of the most commonly applied techniques is the imposition of a low-order autoregressive structure on the errors of a regression model (which may be a static regression apart from the error dynamics). It is well known that this implies a common-factor restriction on the coefficients of a corresponding autoregressive-distributed lag model with white noise errors: that is,

$$y_t = \beta x_t + u_t, \qquad \rho(L) u_t = \varepsilon_t \tag{4.1}$$

is equivalent to

$$\rho(L) y_t = \rho(L) \beta x_t + \varepsilon_t, \tag{4.2}$$

where $\{\varepsilon_t\}$ is a white-noise process, implying a set of restrictions on the coefficients of the regression model (4.2) arising from the common lag polynomial $\rho(L)$. If $\rho(L)$ is of degree k there are k such restrictions; for example, for $k = 2$, $\rho(L) = 1 - \rho_1 L - \rho_2 L^2$ and

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \beta x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \varepsilon_t, \tag{4.3}$$

with $\delta_1 = \rho_1 \beta$ and $\delta_2 = \rho_2 \beta$.

Consider now the effect of using an AR model of error dynamics in this way when the true process contains a moving-average component: that is, the true error process in

(4.1) is instead $\gamma(L)u_t = \theta(L)\varepsilon_t$. The autoregressive-distributed lag (ADL) representation of the model now embodies sets of coefficients on both lagged Y and lagged X, from the approximating AR polynomial $\rho(L)$, which decline geometrically but are non-zero at any finite lag. There is a corresponding (infinite) set of common-factor restrictions. Truncating the representation at any finite lag length k might be expected to perform relatively well as the Hilbert distance to this approximating AR(k) model is smaller. If the distance measure is useful in indicating the order of AR polynomial necessary to model a relation with ARMA errors via an ADL model, there must be a close correspondence between the distance from the ARMA to a given AR, and sample-based indicators of the adequacy of the dynamic specification. The indicator that we use is a standard LM statistic for the null of no autocorrelation from lags 1 to s. The mis-specification considered is the use of an AR(2) error process instead of the true MA(2).

Table 3 reports the results of a simulation experiment designed to check this performance. Using the MA(1) and MA(2) models of Tables 1 and 2b, 5000 replications on samples of size $T = 200$ were generated from the DGP $y_t = \alpha + \beta x_t + u_t$, $\gamma(L)u_t = \theta(L)\varepsilon_t$, with $\alpha = \beta = 1$, $\gamma(L) = I$ and $\theta(L)$ as given in Tables 1, 2b.[4] The innovations $\{\varepsilon_t\}$ have unit variance. The process is modelled with the ADL model corresponding to an AR(2) model of the errors,

$$y_t = \alpha_0 + \sum_{i=1}^{2} \alpha_i y_{t-i} + \sum_{i=1}^{2} \gamma_i x_{t-i} + e_t. \tag{4.4}$$

On each sample, the residuals are tested for autocorrelation up to order ℓ, $\ell = \{1, 2, 12\}$ via an LM test which is asymptotically χ^2_ℓ under the null of no autocorrelation. If the approximation is adequate, then there should be little evidence of residual autocorrelation in these tests. Table 3 gives the mean values of the LM statistics, and ranks both these and

[4]Results on samples of size 1000 are very similar and are therefore not reported.

the corresponding Hilbert and K-L-J measures of the distance between the true process and approximating model; since the approximating model is in every case AR(2), the ranks by distance are in all cases based on the distance to the nearest AR(2).

Both distance measures provide very good *a priori* indicators of the degree to which residual autocorrelation will be detected; that is, they explain the variation in mean LM test statistics very well. The Hilbert distance is especially good; as the order of test increases to measure autocorrelations up to 12, the match by ranks becomes virtually perfect for the Hilbert measure, differing only in the ranking of cases 5 and 6 (ranked 7th and 8th by the Hilbert measure, but 8th and 7th by K-L-J and mean LM). These cases are extremely close, having distance to the nearest AR(2) of 0.189 and 0.188 respectively. The first twelve lagged innovations capture a smaller part of the total variation for process 5 than for 6; k higher than twelve is necessary in the LM test in order to reproduce exactly the Hilbert distance rankings.

[Tables 3, 3b about here]

The use of ADL models to capture dynamics easily through LS regression is commonplace, and is a successful strategy in cases where the error dynamics can be well modelled by a low-order AR. However, where there are MA components with substantial roots, or other components for which the PACF does not approach zero quickly, the Hilbert distance from the DGP of the errors to the AR approximation implicitly used in the ADL specification is a reliable measure of the adequacy of the implicit approximation.

IV. C. Forecasting

Consider next the problem of forecasting a time series process, which may have a moving average component, using a pure autoregression. In this case, a measure of the distance between a given process and the nearest AR(p) will be useful insofar as it gives

an *a priori* indication of the degree to which mean squared error of the forecast is increased by the use of the AR approximation in the place of a model containing MA parts. The process to be forecast is a stationary process $\{y_t\}$, with a Wold representation which we can write as

$$y_{t+1} = f(L)e_{t+1} = f_1(L)e_t + f_0\varepsilon_{t+1}, \tag{4.5}$$

where $e_t = \{\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_1\}'$, and the $\{\varepsilon_t\}$ are white noise. Given a sample of data, we obtain implicitly an estimated lag polynomial $\hat{f}_1(L)$.[5] The one-step-ahead forecast is generated by

$$\hat{y}_{t+1|t} = \hat{f}_1(L)\hat{e}_t, \tag{4.6}$$

where $\hat{y}_{t+1|t}$ indicates a forecast made at time t of the $t+1$ value of Y. The one-step-ahead forecast error is then

$$(\hat{y}_{t+1|t} - y_t) = \hat{f}_1(L)\hat{e}_t - f_1(L)e_t - f_0\varepsilon_{t+1}. \tag{4.7}$$

[Tables 4, 4b about here]

Table 4 gives the mean squared errors of one-step-ahead forecasts made from AR(1), AR(2) and AR(4) models of the MA(2) processes listed in Table 2b, again for $T = 200$ and 5000 replications. Once again, the ordering given by distances of the example processes to the relevant AR approximation matches very well the ordering of the estimated MSE's. In the AR(4) case, the distance and MSE rankings differ only by interchanging cases 4 and 7, which have distance to the nearest AR(4) of 0.128 and 0.139 respectively. Mean squared errors tend to be very close to unity, the correct value for a properly-specified model, for approximation distances of less than 0.1.

[5]For example, if we fit an AR model to the data, $\hat{f}(L)$ represents the projection of the estimated AR polynomial onto past innovations.

Again, both distance measures explain the results well, providing an *a priori* understanding of the MA or ARMA parameter values that allow a good approximation to be made with an AR of given order. The Hilbert distance seems again to have some advantage. For the AR(4) case, the Hilbert ranking differs from that of the forecast errors only for cases 4 and 7 (ranked 6th and 5th, respectively, rather than 5th and 6th). The K-L-J ranking is similar to that of the Hilbert distance, but makes an additional interchange relative to the ranking of cases by forecast error, in cases 2 and 3.

IV. D. Evaluation of information criteria

Sample-based selection of appropriate lag length (or, more generally, model order) is often based on information criteria such as those of Akaike, Schwarz, and others; see [21] and [23] for recent reviews. In the context of problems for which the DGP is a special case of more general estimated models, we can investigate these criteria by simulation, preferring those which tend to yield lag lengths close to the optimal values. Where the model is an approximation, however, it may be unclear what the best lag length is even in a constructed example, so that evaluation of the criteria in cases such as that of AR models which are being used to approximate more general processes cannot proceed.

However, using a distance measure of the difference between DGP and AR approximation, we can proceed as in section IV.A to an answer to the question of what the optimal approximating model order is, given a DGP and sample size. From this it is possible to evaluate the information criteria, by examining the degree to which the typical selected lag length differs from the optimum. This section provides a brief example of such an exercise, using the AIC, BIC, Schwarz and FPE criteria.[6]

[6]For this linear regression problem the criteria can be reduced to the following expressions in the sample size, T, number of autoregressive terms, k, and sum of squared residuals, $\hat{\varepsilon}'\hat{\varepsilon}$:

For the data generation processes and sample sizes in section IV.A, we compute the average lag lengths selected by each of these criteria, in 2500 simulated samples. The results are recorded in Table 5, along with the optimal approximating lag lengths from Figures 1 to 3. Where the objective function is nearly flat near the optimum lag length, we report a range of optimal values (e.g., 8–10 for $T = 1000$ and $\theta = 0.9$). The set of lag lengths considered ranged from 1 to 20; with even larger values included, averages for the AIC would rise slightly.

[Table 5 about here]

The BIC and Schwarz criteria, which are very similar and closely related, produce very good results. The AIC, as has been observed in contexts where approximation and mis-specification play no role, over-parameterizes dramatically. The FPE falls in between, over-parameterizing consistently, but less substantially than the AIC.

V. Concluding remarks

There are many circumstances in which it is convenient to approximate an ARMA process by a pure AR(p) process. But while the technique is widely used, often implicitly, there are relatively few results concerning the order of autoregression necessary to provide a good approximation. This paper addresses the question of the quality of an approximation using measures of the distance between processes, primarily the Hilbert distance. By minimizing this distance from a process to a class of processes, we are able to find the closest process of given order in the target class. The results offer a general contribution to understanding of the relations between ARMA processes, of the gains available from more elaborate modelling, and of the use of autoregressive approximations in various applied

---

AIC: $\ln(\hat{\varepsilon}'\hat{\varepsilon}/T) + 2k/T$; BIC: $\ln(\hat{\varepsilon}'\hat{\varepsilon}/T) + k\ln(T)/T$; Schwarz: $\ln(\hat{\varepsilon}'\hat{\varepsilon}/(T-k)) + k\ln(T)/T$; FPE: $\frac{(T+k)}{(T-k)}(\hat{\varepsilon}'\hat{\varepsilon}/(T-k))$.

problems including the traditional problem of choice of order.

Acknowledgments

The authors thank Alfred Haug, David Hendry, Aman Ullah and seminar participants at Amsterdam, Erasmus, and Oxford universities, CORE, the Canadian Econometric Study Group and Société canadienne des sciences économiques for valuable comments. The Fonds pour la Formation de chercheurs et l'aide à la recherche (Quebec) and the Social Sciences and Humanities Research Council of Canada provided financial support for this research.

References

[1] W.A. Fuller. Introduction to Statistical Time Series. New York: Wiley, 1976.

[2] J. Durbin. Efficient estimation of parameters in moving-average models. Biometrika 46:306-316, 1959.

[3] J. Durbin. The fitting of time series models. Review of the International Statistical Institute 28:233-243, 1960.

[4] E.J. Hannan, J. Rissanen. Recursive estimation of mixed autoregressive-moving average order. Biometrika 69:81-94, 1982.

[5] P. Saikkonen. Asymptotic properties of some preliminary estimators for autoregressive moving average time series models. Journal of Time Series Analysis 7:133-155, 1986.

[6] S. Koreisha, T. Pukkila. A generalized least squares approach for estimation of autoregressive moving average models. Journal of Time Series Analysis 11:139-151, 1990.

[7] J.W. Galbraith, V. Zinde-Walsh. A simple, non-iterative estimator for moving-average models. Biometrika 81:143-155, 1994.

[8] J.W. Galbraith, V. Zinde-Walsh. Simple estimation and identification techniques for general ARMA models. Biometrika 84:685-696, 1997.

[9] H. Akaike. Power spectrum estimation through autoregressive model fitting. Annals of the Institute of Statistical Mathematics 21:407-419, 1969.

[10] K.N. Berk. Consistent autoregressive spectral estimates. Annals of Statistics 2:489-502, 1974.

[11] S.E. Said, D.A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71:599-607, 1984.

[12] J.W. Galbraith, V. Zinde-Walsh. On the distributions of Augmented Dickey-

Fuller statistics in processes with moving average components. Journal of Econometrics 93:25-47, 1999.

[13] E. Maasoumi. A compendium to information theory in economics and econometrics. Econometric Reviews 12:137-181, 1993.

[14] A. Ullah. Entropy, divergence and distance measures with econometric applications. Journal of Statistical Planning and Inference 49:137-162, 1996.

[15] Y.A. Rozanov. Stationary Random Process. San Francisco: Holden-Day, 1967.

[16] M.B. Priestley. Spectral Analysis and Time Series. London: Academic Press, 1981.

[17] V. Zinde-Walsh. The consequences of mis-specification in time series processes. Economics Letters 32:237-241, 1990.

[18] J. Burbea, C.R. Rao. Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. Journal of Multivariate Analysis 12:575-596, 1982.

[19] E. Parzen. Autoregressive spectral estimation. In: D.R. Brillinger, P.R. Krishnaiah, eds., Handbook of Statistics, v. 3. Amsterdam: North-Holland, 1983, pp 221-247.

[20] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling. Numerical Recipes: the Art of Scientific Computing. Cambridge: Cambridge University Press, 1986.

[21] B.S. Choi. ARMA Model Identification. New York: Springer-Verlag, 1992.

[22] D.F. Hendry. Dynamic Econometrics. Oxford: Oxford University Press, 1995.

[23] J.A. Mills, K. Prasad. A comparison of model selection criteria. Econometric Reviews 11:201-233, 1992.

Table 1

Approximation distances:[7]

Distance from MA(1) process to nearest AR(p)

Root	Order, p, of approximating AR process					
θ	*0*	*1*	*2*	*4*	*8*	*12*
.999	1.081	0.570	0.366	0.199	9.37×10^{-2}	5.70×10^{-2}
.99	1.071	0.563	0.360	0.195	9.04×10^{-2}	5.43×10^{-2}
.95	1.023	0.530	0.335	0.176	7.64×10^{-2}	4.29×10^{-2}
.90	0.964	0.490	0.303	0.152	6.02×10^{-2}	3.04×10^{-2}
.70	0.734	0.335	0.185	7.16×10^{-2}	1.51×10^{-2}	3.54×10^{-3}
.50	0.514	0.196	8.75×10^{-2}	2.05×10^{-2}	1.27×10^{-3}	8.09×10^{-5}
.30	0.303	8.05×10^{-2}	2.36×10^{-2}	2.11×10^{-3}	1.72×10^{-5}	1.46×10^{-7}
.10	0.100	9.85×10^{-3}	9.85×10^{-4}	9.86×10^{-6}	9.86×10^{-10}	1.00×10^{-13}
.05	0.050	2.49×10^{-3}	1.25×10^{-4}	3.11×10^{-7}	1.96×10^{-12}	1.22×10^{-18}
.01	0.010	1.00×10^{-4}	1.00×10^{-6}	1.00×10^{-10}	1.00×10^{-18}	1.00×10^{-26}

[7]In Tables 1 and 2a, the column headed "*0* " gives the distance to the white noise process having the same variance. Results in Table 1 are unaffected by multiplying the moving-average parameter by -1.

Table 2a

Approximation distances:[8]

Distance from MA(2) process to nearest AR(p)

	Order, p, of approximating AR process					
Case	*0*	*1*	*2*	*4*	*8*	*12*
1	2.605	1.326	0.792	0.376	0.137	6.57×10^{-2}
2	2.368	1.178	0.683	0.299	8.34×10^{-2}	2.78×10^{-2}
3	1.095	0.569	0.362	0.194	8.95×10^{-2}	5.36×10^{-2}
4	1.785	0.818	0.421	0.128	5.58×10^{-2}	2.03×10^{-2}
5	1.225	0.477	0.189	8.47×10^{-2}	2.05×10^{-2}	5.08×10^{-3}
6	0.990	0.404	0.188	4.85×10^{-2}	3.55×10^{-3}	2.41×10^{-4}
7	0.604	0.446	0.259	0.139	5.28×10^{-2}	2.41×10^{-2}
8	1.680	0.792	0.436	0.171	4.01×10^{-2}	1.15×10^{-2}
9	0.142	0.108	2.19×10^{-2}	3.39×10^{-3}	7.03×10^{-5}	1.36×10^{-6}
10	0.457	0.305	0.158	6.60×10^{-2}	1.38×10^{-2}	3.09×10^{-3}
11	0.766	0.245	6.87×10^{-2}	1.14×10^{-2}	2.52×10^{-4}	1.29×10^{-5}
12	0.0283	1.96×10^{-2}	8.89×10^{-4}	2.47×10^{-5}	1.50×10^{-8}	8.08×10^{-12}

[8]The case numbers refer to Table 2b, where the processes are described.

Table 2b

Features of MA(2) processes used in Table 2a

	MA parameters		*Real parts*		*Imaginary parts*		*Moduli*	
Case	θ_1	θ_2						
1	-1.96	0.98	0.980	0.980	0.140	-0.140	0.990	0.990
2	-1.80	0.90	0.900	0.900	0.300	-0.300	0.949	0.949
3	-1.01	0.0198	0.990	0.020	0.00	0.00	0.990	0.020
4	-1.40	0.70	0.700	0.700	0.458	-0.458	0.837	0.837
5	1.00	0.50	-0.500	-0.500	0.500	-0.500	0.707	0.707
6	0.90	0.20	-0.500	-0.400	0.00	0.00	0.500	0.400
7	-0.50	-0.30	0.852	-0.352	0.00	0.00	0.852	0.352
8	-1.40	0.49	0.700	0.700	0.00	0.00	0.700	0.700
9	0.10	-0.10	-0.370	0.270	0.00	0.00	0.370	0.270
10	0.40	-0.20	-0.690	0.290	0.00	0.00	0.690	0.290
11	-0.70	0.20	0.350	0.350	0.278	-0.278	0.447	0.447
12	0.020	0.02	-0.010	-0.010	0.141	-0.141	0.141	0.141

Table 3

LM tests for residual autocorrelation:
MA errors modelled by AR approximation
T = 200

Case	θ_1	θ_2	$\ell = 1$	$\ell = 2$	$\ell = 12$
1	-1.96	0.98	31.64	51.36	95.95
2	-1.80	0.90	34.39	55.86	92.77
3	-1.01	0.0198	12.73	21.12	46.11
4	-1.40	0.70	28.58	39.66	53.69
5	1.00	0.50	3.882	5.555	20.03
6	0.90	0.20	8.507	12.55	22.85
7	-0.50	-0.30	7.529	13.57	31.30
8	-1.40	0.49	25.22	39.76	63.77
9	0.10	-0.10	1.055	2.138	12.31
10	0.40	-0.20	4.308	7.556	18.95
11	-0.70	0.20	1.999	3.106	13.14
12	0.020	0.02	1.026	2.093	12.21

Table 3b

Cases ranked by approximation distance and LM test
(rank of given case by: K-L-J distance,Hilbert distance,LM statistic)
T = 200

Case	θ_1	θ_2	$\ell = 1$	$\ell = 2$	$\ell = 12$
1	-1.96	0.98	(1, 1 ,2)	(1, 1 ,2)	(1, 1 ,1)
2	-1.80	0.90	(2, 2 ,1)	(2, 2 ,1)	(2, 2 ,2)
3	-1.01	0.0198	(3, 5 ,5)	(3, 5 ,5)	(3, 5 ,5)
4	-1.40	0.70	(5, 4 ,3)	(5, 4 ,4)	(5, 4 ,4)
5	1.00	0.50	(8, 7 ,9)	(8, 7 ,9)	(8, 7 ,8)
6	0.90	0.20	(7, 8 ,6)	(7, 8 ,7)	(7, 8 ,7)
7	-0.50	-0.30	(6, 6 ,7)	(6, 6 ,6)	(6, 6 ,6)
8	-1.40	0.49	(4, 3 ,4)	(4, 3 ,3)	(4, 3 ,3)
9	0.10	-0.10	(11,11,11)	(11,11,11)	(11,11,11)
10	0.40	-0.20	(9, 9 ,8)	(9, 9 ,8)	(9, 9 ,9)
11	-0.70	0.20	(10,10,10)	(10,10,10)	(10,10,10)
12	0.020	0.02	(12,12,12)	(12,12,12)	(12,12,12)

Table 4

MSE's of one-step-ahead forecasts:
MA processes modelled by AR approximation
T = 200

Case	θ_1	θ_2	AR(1)	AR(2)	AR(4)
1	-1.96	0.98	3.278	2.460	1.821
2	-1.80	0.90	2.792	2.058	1.476
3	-1.01	0.0198	1.514	1.340	1.219
4	-1.40	0.70	1.839	1.357	1.098
5	1.00	0.50	1.268	1.077	1.065
6	0.90	0.20	1.236	1.088	1.039
7	-0.50	-0.30	1.261	1.141	1.081
8	-1.40	0.49	1.872	1.457	1.177
9	0.10	-0.10	1.022	1.017	1.032
10	0.40	-0.20	1.125	1.055	1.041
11	-0.70	0.20	1.081	1.023	1.032
12	0.020	0.02	1.010	1.017	1.032

Table 4b

Cases ranked by approximation distance and one-step MSE
(rank of given case by: K-L-J distance,Hilbert distance,MSE)
T = 200

Case	θ_1	θ_2	AR(1)	AR(2)	AR(4)
1	-1.96	0.98	(1, 1 ,1)	(1, 1 ,1)	(1, 1 ,1)
2	-1.80	0.90	(2, 2 ,2)	(2, 2 ,2)	(3, 2 ,2)
3	-1.01	0.0198	(3, 5 ,5)	(3, 5 ,5)	(2, 3 ,3)
4	-1.40	0.70	(5, 3 ,4)	(5, 4 ,4)	(6, 6 ,5)
5	1.00	0.50	(8, 6 ,6)	(8, 7 ,8)	(7, 7 ,7)
6	0.90	0.20	(7, 8 ,8)	(7, 8 ,7)	(9, 9 ,9)
7	-0.50	-0.30	(6, 7 ,7)	(6, 6 ,6)	(5, 5 ,6)
8	-1.40	0.49	(4, 4 ,3)	(4, 3 ,3)	(4, 4 ,4)
9	0.10	-0.10	(11,11,11)	(11,11,11)	(11,11,11)
10	0.40	-0.20	(9, 9 ,9)	(9, 9 ,9)	(8, 8 ,8)
11	-0.70	0.20	(10,10,10)	(10,10,10)	(10,10,10)
12	0.020	0.02	(12,12,12)	(12,12,12)	(12,12,12)

Table 5

Estimated optimal AR order vs.
Mean selected order, various criteria

	Case	*Optimal*	*Average selected order*			
θ	T	order	AIC	BIC	Schwarz	FPE
0.1	200	1	12.1	1.14	1.06	3.47
0.1	1000	1	10.1	1.05	1.02	2.89
0.5	200	2–3	12.7	2.20	1.94	4.91
0.5	1000	3–4	11.4	2.96	2.80	5.49
0.9	200	4–5	16.2	5.98	5.09	11.1
0.9	1000	8–10	17.8	9.91	9.16	15.8

Figure 1
Approximation distance vs. AR order
MA(1) parameter = 0.90

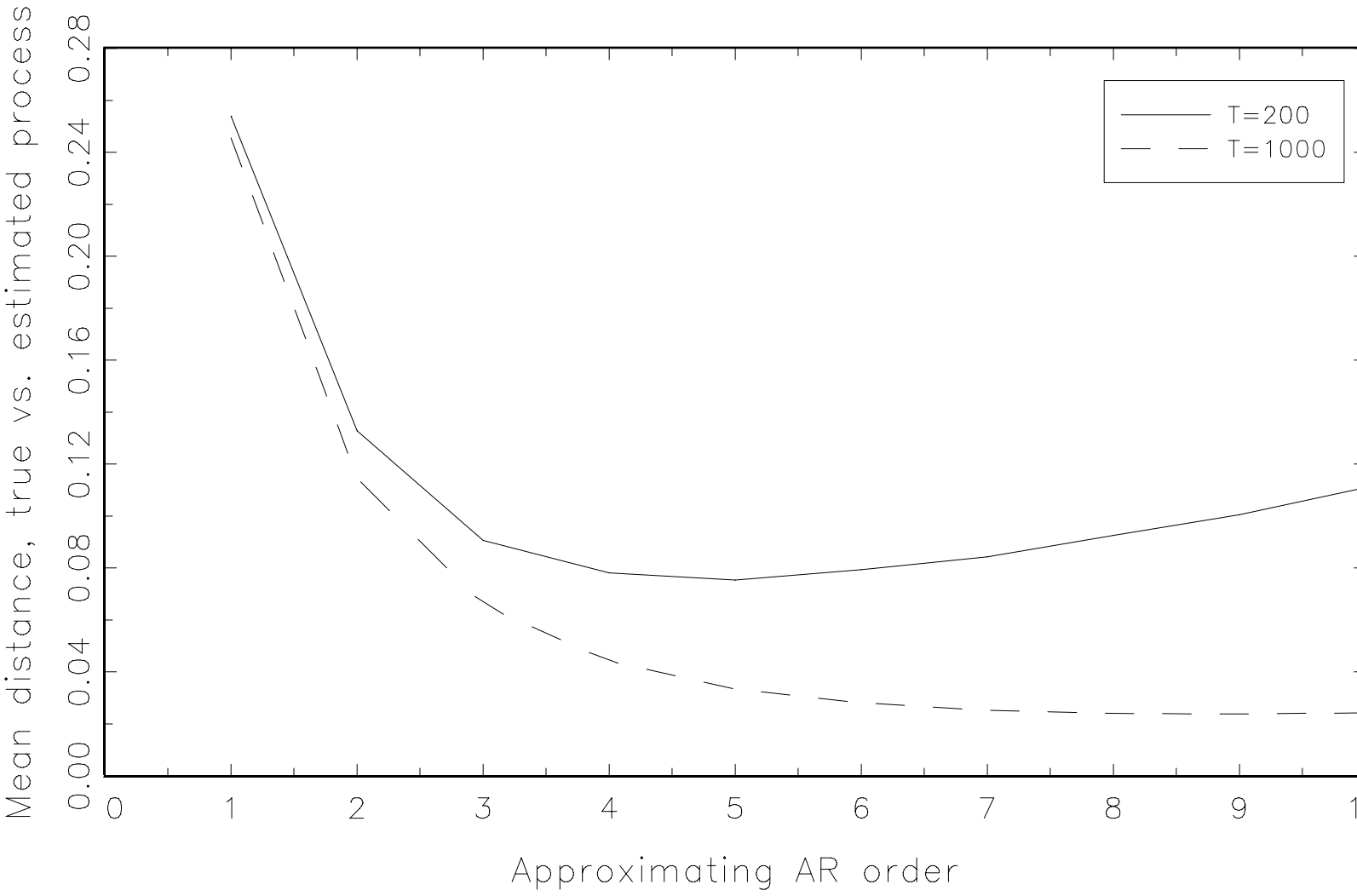

Figure 2
Approximation distance vs. AR order
MA(1) parameter = 0.50

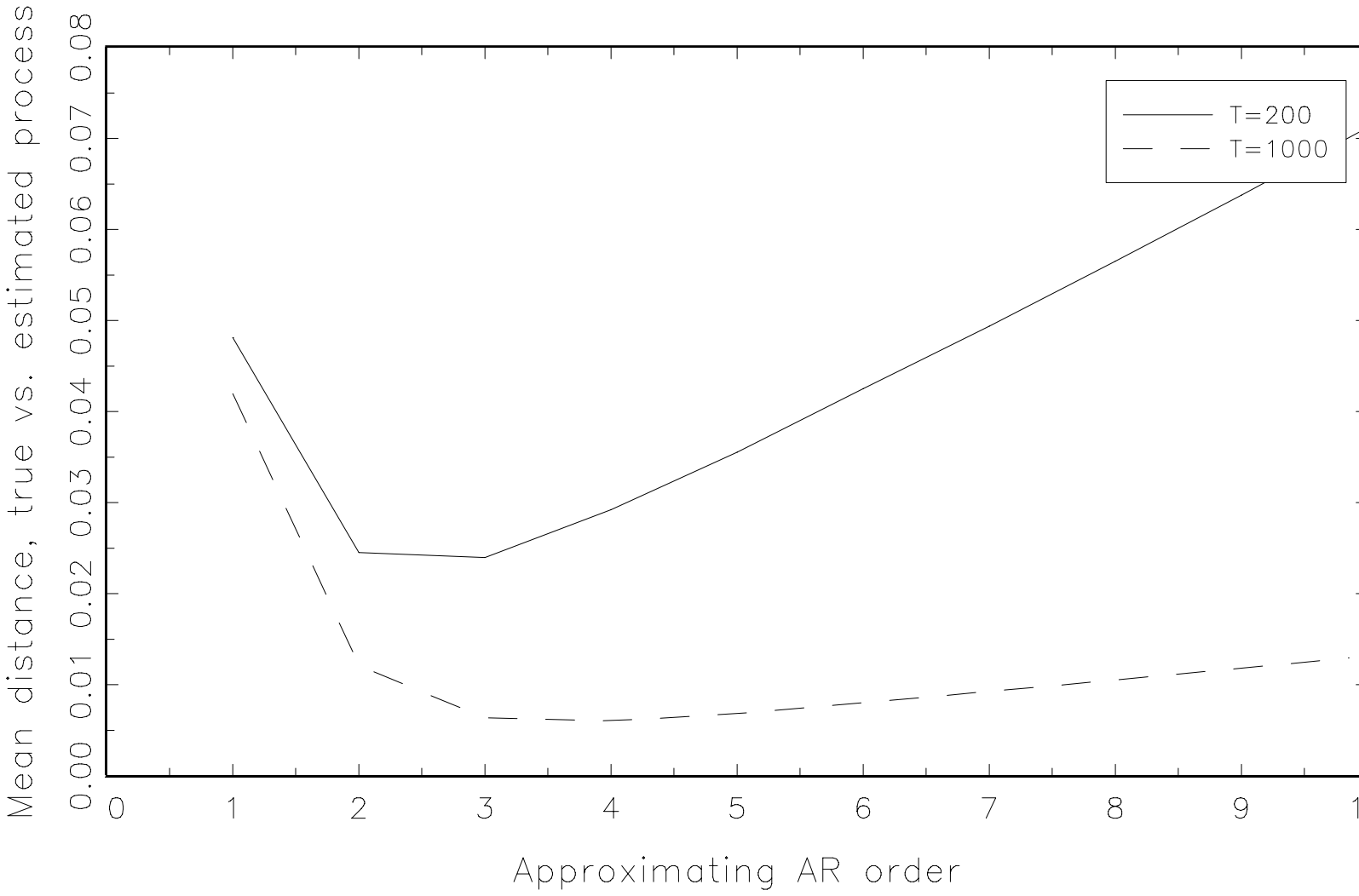

Figure 3
Approximation distance vs. AR order
MA(1) parameter = 0.10