A BAYESIAN ANALYSIS FOR ORDER DETERMINATION OF AUTOREGRESSIVE MOVING AVERAGE TIME SERIES MODEL.

SUNDARABALAN. S* ARUMUGAM. P[1] & SARANRAJ. R[2]
*Research scholar Department of Statistics, ManonmaniamSundaranar University,Tamilnadu, India.
[1]Assistant Professor, Department of Statistics, Annamali University.
[2]Assistant Professor, Department of Statistics, St.Joseph's College of Arts & Science, Cuddalore.

Abstract

In Time Series analysis, the order of various Time series models plays an important role in studying model selection criterion. Hannan and Quinn(1979) have been studied the order determination of an autoregressive moving average through classical approach. In this paper, an attempt is made to study the order determination of autoregressive moving average time series model by employing Bayesian methodology.

key words: Autoregressive moving average, Order determination, Bayes theorem, Model selection.

1. Introduction

The most popular approach to identify the orders of ARMA (p,q) models is developed by Box and Jenkins (1970). There are situations where the time series may be represented as a mix of both AR and MA models referred as ARMA (p,q) The Bayesian identification of time series is being developed and the Bayesian literature devoted to the analysis of ARMA models is sparse. Diaz and Farah (1981) developed a direct Bayesian method to identify the autoregressive moving average models.

Bayesian inference holds a distinct advantage over so-called classical statistics in non-standard problems where concepts such as sufficiency or completeness do not apply. That advantage is that the program is unchanged: the prior together with the likelihood produce the posterior. A disadvantage is that conjugate families are not available and so the Bayes Theorem must be used numerically, for which approximation and numerical integration techniques are required.

Autoregressive moving average (ARMA) Time series models are quite nonstandard, even if the usual assumption of normality is retained. The number of parameters, reflected in the order of the model (p,q), is undetermined. Given (p,q), the parameter space must then be constrained for identifiability reasons.

Classical analysis of ARMA modelsmust rely on asymptotic behavior: consistency, asymptotic normality, and efficiency.

Bayesian statistics are little affected by sample size. Moreover, the asymptotic in a Bayesian analysis are substantially the same as the classical. But in small sample problems, the choices of prior distribution and loss function do influence the consequential decision. In cases such as these, the effort in expressing these two are well rewarded.

Bayesian analyses of non-standard problems are commonly believed to suffer from the profound defect of computational intractability. This chapter finds that belief directly, proving that a Bayesian analysis of ARMA models can be done. The proof is a computational method that has been successfully implemented and that can be extended in its sophistication. Thus most of the effort lies in solving the challenging computational problems that arise.

In section 4, the parameter structure is laid out and the form of the prior distribution is designed. Secondly, the dual tasks of computing and expressing numerically the posterior distribution are discussed in section 5. The inference novel to this Bayesian analysis is a method of selecting the order of the model (p,q). This methodology is demonstrated using as examples two well known series in section 6. Classical methods of estimation and model selection are discussed in section 3. Explanation of the probability structure of the ARMA model follows in section 2.

2. Autoregressive Moving average model

The autoregressive moving average (ARMA)process$\{x_t\}$ of order (p,q) is defined by the stochastic difference equation

$$(X_t - \mu) - \alpha_1(X_{t-1} - \mu) - \cdots - \alpha_p(X_{t-p} - \mu) = e_t - \beta_1 e_{t-1} - \cdots - \beta_q e_{t-q} \ldots (1)$$

where $e_t, t = \cdots, -3, -2, -1, 0, 1, 2, 3, \ldots$ are i.i.d. Normal $(0, \sigma_e^2)$ random variables. A finite segment $x = (x_1, x_2, \ldots x_n)^T$ is observed, which has a multivariate normal distribution with mean vector $\mu 1_N$, and covariance matrix $\sigma_e^2 A_N$ (here 1_K, is k by 1 with all entries equal to one), denoted by

$$X \sim N(\mu 1_N, \sigma_e^2 A_N) \quad \ldots (2)$$

The matrix A_N, is defined by

$$(A_N)_{ij} = \mathrm{cov}(x_i, x_j) = \rho(i-j) = \rho(|i-j|) \ldots (3)$$

where the covariance function ρ depends upon $\alpha_1, \alpha_2, \ldots \alpha_p, \beta_1, \beta_2 \ldots \beta q$ and characterizes the time series process $\{x_t\}$. The n periods of the process to be forecast, $X_F = \left(X_{N+1,} X_{N+2,} \ldots\ldots X_{N+n}\right)^T$, when conditioned upon the observed x, then has an n-dimensional multivariate normal distribution with mean vector

$$\mu 1_n + A_{21} A_N^{-1} (X - \mu 1_N) \equiv \mu a + b \ldots (4)$$

and covariance matrix $\sigma_e^2 A^*_{nN}$ where

$$A^*_{N+n} = \begin{bmatrix} A_N & A_{12} \\ A_{21} & A_n \end{bmatrix}^N_n,$$

and

$$A^*_{nN} = A_n - A_{21} A_N^{-1} A_{12}$$

Two conditions must be enforced on the autoregressive moving average parameters, $\Theta \equiv \left(\alpha_1, \alpha_2, \ldots . \alpha_p, \quad \beta_1, \beta_2 \ldots, \beta_q, \right)^T$, of this model. The stationarity condition states that the roots of the polynomial equation $\sum \alpha_i w^i = 0$ must lie outside the unit circle in the complex plane $\left(\alpha_0 = -1\right)$ Otherwise, the process is explosive and an indefinite past for it cannot exist. Secondly, the identifiability condition states that the roots of the equation $\sum \beta_i w^i = 0$ lie on or outside the unit circle. These identifiability conditions here enforce a unique parameterization of the model in terms of μ, σ_e^2, and the ARMA parameters Θ.

A slightly different probability model is sometimes posed where starting values $x_0, \ldots x_{1-p}$ are introduced as parameters or where starting conditions $a_{p+1-q} = \cdots = a_p = 0$ are enforced. Conceptually, this would allow for models with explosive autoregressive moving average behavior, but it also can lead to the identifiability problems discussed by Pagano (1973). The big advantage is a simpler probability structure which permits evaluation of the likelihood in O (N) operations. This is the motivation for the 'conditional

least squares' estimation discussed by Ansley and Newbold (1980). A drawback is dealing with the starting values as nuisance parameters. The computational advantage was eliminated by Ansley (1979) who showed that the exact likelihood of the model in eq. (1) can be evaluated in O (N) also.

Lastly, ARMA models exhibit what might be called near-nonidentifiability. The parameterization just described is truly unique if the requisite conditions are enforced. But for finite samples, quite different parameter values can produce very similar distributions. In terms of moments, the variance and first three autocovariances of an AR(1) process with α= 0.4 and σ^2= l are 1.12, 0.32, 0.13, and 0.05; for MA(1) with β= -0.5, these values are 1.07, 0.2, 0 and 0 Only with many observations can these be distinguished. From the likelihood viewpoint, the difference in expected log likelihood for these two parameter values is -0.176 when the AR model is true and N =50, while the variance of the log-likelihood is 25 when evaluated at the true. Near-non-identifiability becomes more of a problem with the more flexible mixed and higher order models.

3. Model selection and estimation

The problem of determining (p,q) has led to three different approaches. A fourth, a Bayesian procedure due to Akaike, will be discussed later. The first approach is to extend the standard regression t-test or F-test to the autoregressive moving average model. For pure AR models, rewrite (1) as

$$x_t = x_{t-1}\alpha_1 + x_{t-2}\alpha_2 + \ldots + x_{t-p}\alpha_p + e_t \quad \ldots (5)$$

and regress x_t, on the lagged values (x_{t-1} x_{t-2} x_{t-p}) to obtain estimates of α . The standard t-test of $\alpha_1 = 0$ or F-test of $\alpha_k = \alpha_{k+1} = \ldots = \alpha_t = 0$ can then be used to determine the true value of p. This procedure is discussed in detail by Anderson (1970).

A second procedure is Akaike's (1974) MAICE criterion: find (p,q) to minimize

AIC = - 2 log (maximum likelihood) + $2(p+q)$.

This procedure can be viewed as an adjustment for the number of parameters in a likelihood ratio test, or as an analogue to the F-test described above, with AIC ('adjusted information criterion')

corresponding to an adjusted sum of squares. This procedure has been modified for consistency [see Hannan and Quinn (1979)].

A third approach relies on a family of goodness-of-fit procedures. Assuming the correct (p,q) and values of Θ, the asymptotic distribution of the residual correlations have a zero mean. For model selection, the informal approach of Box and Jenkins (1976) advocates selecting a tentative model (p,q), estimating its parameters, and 'diagnostic checking: performing a test for goodness-of-lit. Upon rejection, the process is repeated with a different model. One might formalize this to selecting the model (p,q) as the one that performs best on a goodness-of-fit test, while adjusting for the number of parameters, as with AIC. For a given (p,q) several proposals have been made for estimating ARMA parameters. Estimators based on moments, such as solving the Yule-Walker equations [see Fuller (1978)] and/or regression have an obvious appeal because of the directness of the computation. Ansley and Newbold entertain three implicit estimators in their Monte-Carlo study: maximum likelihood, 'exact least squares', and 'conditional least squares'. Exact least squares ignores the effect of the ARMA parameters on the likelihood through the determinant of A_N. Conditional least squares uses the model with starting values $a_{p+1-q} = \ldots = a_p = 0$ which requires much less computational effort. Although no procedure receives wide acceptance, Ansley and Newbold (1980) strongly recommend maximum likelihood as a result of their small sample (N = 50)simulation study.

Finally, the work of Akaike (1979) in extending (in pure AR models) his MAICE procedure in a Bayesian fashion must not be overlooked. He first relates the modification of the term 2p in AIC to αp to placing a geometric type prior on the order of the model. He then introduces a loss function based on prediction error to be minimized to find the best order of theautoregression. Although Akaike uses the term 'full Bayesian', it differs substantially from the Bayesian attitude of this chapter. What is used as a likelihood is not averaged over the subsidiary parameters, nor approximated by the maximum, but an approximation based on solving the Yule-Walker equations using sample autocorrelations.

4. The structure of the prior

Three things are required to perform a Bayesian analysis. The first requirement is to correct the structure for the parameterization of the problem. In time series, such a structure can be found in the ARMA

model. Implicit in this parameter structure is the form of the joint prior distribution on all of the model parameters: $p, q, \alpha, \beta, \mu, \sigma_e^2$Secondly, a vehicle is required toexpress information in the posterior distribution as completely as possible, emphasized by Box and Tiao (1973, p.14). Finally, a method is needed to compute the necessary posterior densities and moments. The goal here is as much detail as is computationally feasible, whether in moments or in marginal densities of parameters or forecasts. Most of the computational effort is in numerical integration, and the intention here is to allow the greatest flexibility in specifying the prior, so that minor changes do not require extensive recomputation.

As previously mentioned, the ARMA model provides a structure for expression of the joint prior distribution on the parameters. This is done conditionally, beginning with the coarsest level of parameterization, the order of the model, (p,q). The prior distribution on M=m is given as probabilities,

Pr(M = m) = Pr (x arises from an ARMA (p_m,q_m) process = p_m,on the discrete set of pairs of non-negative integers. Here m may be considered as in Monahan(1982) such that ,

$m = (q_m + 1) + (p_m + q_m)(p_m + q_m + 1)/2$, asin the following table:

m	(p, q)	m	(p,q)
1	(1,0)	4	(2,0)
2	(0,1)	5	(0,2)
3	(1,1)	6	(2,2)

Conditional on M = m, the order of the model, a prior distribution on the ARMA parameters, $\Theta = (\alpha_1, \alpha_2, ...\alpha_{pm}, \ \beta_1, \beta_2 ..., \beta_{qm},)^T$, is specified;π(Θ|M = m) being a density on $R^{pm+qm}$with support on$c_{p\times}c_q$.

where

$$C_K \equiv \left\{ y: \begin{array}{l} \text{alltherootsof } 1 - \sum y_i w^i = 0 \\ \text{exceedoneinabsolutevalue} \end{array} \right\}$$

y' exceed one in absolute value '

This restriction expresses no prior information but only enforces a unique parameterization by insuring that α and βsatisfy the stationarity and identifiability conditions. The remaining parameters are now μ, the mean of the process, and σ_e^2, the disturbance variance. A convenient reparameterization is $r = 1/\sigma_e^2$with r called the disturbance precision. While there is no conjugate prior for Θ, using the standard normal-gamma conjugate family for μ and r is quite useful,and

$$(\mu|r,\Theta,M) \sim N(\gamma,(\tau r)^{-1}) \ldots (6)$$

$$(r|\Theta,M) \sim gamma(\alpha,\beta) \ldots (7)$$

Notice that the parameters of this joint prior may depend on Θ and M without causing further complications, thus prior information on, say, the series variance can be reflected in the prior on r.

With the prior now specified, recall the probability distribution for the data given in section 6.2,

$$(x|M,\Theta,\mu,r) \sim N(\mu 1, r^{-1}A_N) \ldots (8)$$

Two useful marginal posterior distributions, both multivariate t, arevaluable.

$$(x|M,\Theta) \sim t_{N,2\alpha}(\gamma 1_N, (\alpha/\beta)[A_N + \tau^{-1}1_N 1_N^T]^{-1}) \ldots (9)$$

and

$$(x_F|x,M,\Theta) \sim t_{n,2\alpha+N}\left(\gamma^* a+b, (2\alpha+N/\beta^*)\left[A_{n\,N}^* + \tau^{-*}aa^T\right]^{-1}\right) \ldots (10)$$

where

$$\gamma^* = (\gamma\tau + 1^T A_N^{-1}1)/\tau^*$$

$$\tau^* = 1/\tau^{-*} = \tau + 1^T A_N^{-1}1$$

$$\beta^* = \beta + (z-\gamma 1)^T (A_N + \tau^{-1}11^T)^{-1}(z-\gamma 1),$$

and the vectors a and b are defined in (6.4).

The remainder of the Bayesian analysis now follows easily, beginning with the marginal of x,

$$f(x|M) = \int \pi(\Theta|M) f(x|\Theta, M) d\Theta \quad ...(11)$$

where $f(x|\Theta, M)$ is the multivariate t density described in (6.9). This integration must be done numerically and will be described later. The posteriors now follow automatically:

$$\pi(\Theta|M, x) = \pi(\Theta|M) f(x|\Theta, M) / f(x|M) \quad ...(12)$$

thus $f(x|M)$ is the normalizing constant in the posterior density of the ARMA parameters. For the posterior probability of M =m, $f(x|M=m)$ serves as a Bayes Factor,

$$p(M|x) = p(M) f(x|M) \Big/ \sum_j p(M=j) f(x|M=j) \quad ...(13)$$

And the posterior density of the forecasts can be expressed by

$$f(x_F|x) = \sum_j p(M=j|x) \int f(x_F|x, \Theta, M=j) \pi(\Theta|x, M=j) d\Theta, \quad ...(14)$$

where $f(x_F|x, \Theta, M)$ is the multivariate t density in (6.10). The relevant posterior distributions were derived in Monahan (1980a).

5. Computation and expression of the posterior

The thrust of this chapter is the demonstration of the feasibility of a Bayesian analysis of ARMA time series models. The Bayesian approach is nearly always straightforward: once the model is specified, giving the likelihood, the prior then determines the posterior. In situations where conjugate families are available, this procedure is relatively simple. However, in this non-standard problem, three difficult computational problems arise: computing $f(x_F|\Theta, M)$ efficiently, numerical integration of (11) to obtain $f(x|M)$, and expressing as much of the posterior as computationally practical. The feasibility of a Bayesian procedure hinges on the solution of these problems.

The first computational problem involves computing $f\left(x|\Theta,M\right)$ which serves as the likelihood. Ansley's (1978, 1979) method for computing the exact likelihood in ARMA models cannot be directly applied since it does not admit a mean parameter and also because $f\left(x|\Theta,M\right)$ is not the likelihood function for the standard ARMA model. However, the principle of Ansley's method can be adapted to evaluate $f\left(x|\Theta,M\right)$ quickly.

Essentially, only bilinear forms need to be computed, since

$$f\left(x|\Theta,M\right)\propto\left[1+\frac{1}{2\beta}(x-\gamma 1)^T(A_N+\tau^{-1}11^T)^{-1}(x-\gamma 1)\right]^{-(\alpha+N/2)}$$

The technique hinges on following factorizations:

$$B_{m,N}A_N\left(B_{m,N}\right)^T=LDL^T=\begin{pmatrix}A_m & D_1^T\\ D_1 & C\end{pmatrix}\begin{matrix}\mathrm{m}=\max(\mathrm{p},\mathrm{q})\\ \mathrm{N}-\mathrm{m}\end{matrix}$$

Here $B_{m,N}$ is a triangular matrix composed of I_m in its upper left sub matrix and other rows of $-\alpha_p,\ldots,-\alpha_1,1$ with the ones aligned on the diagonal and with zeros elsewhere. The matrix D,is mostly zeros and the entire far right-hand side above has a bandwidth of 2m+1. Hence the Cholesky factorization without square roots, LDL^T, results in a diagonal matrix D and a unit lower triangular matrix L with bandwidth m+1. The bilinear forms are then computed from

$$x^TA_N^{-1}y=x^TB_{m,N}^{-T}L^{-T}D^{-1}L^{-1}B_{m,N}^{-1}y$$

$$=\left(L^{-1}B_{m,N}^{-1}x\right)^T D^{-1}\left(L^{-1}B_{m,N}^{-1}y\right)$$

Computation of the forecast covariance matrix, though not a bilinear form, utilizes these same factorizations. Monahan (1980b) gives the many details and a listing of the code (about 500 lines). The important point of the solution of this problem is that the effort, in both time and space, for computing $f\left(x|\Theta,M\right)$ is roughly proportional to N.

As previously stated, conjugate families are available for the parameters p,q and r. But none are available forΘ, hence its posterior, $\pi(\Theta \mid M, x)$, which requires normalization by $f(x \mid M)$ given by (6.1l),

$$f(x \mid M) = \int \pi(\Theta / M) f(x \mid \Theta, M) d\Theta,$$

Must be computed numerically. The effort required in this numerical integration depends substantially on two criteria: smoothness of the integrand and the number of dimensions.

Smoothness is not a major concern. A smooth prior would be expected and the likelihood, $f(x \mid \Theta, M)$ should resemble a normal density even for small sample sizes and near-non-identifiability of the parametersΘ. Notice also that the likelihood vanishes on the boundary of the region of integration, $C_p \times C_q$

The dimensionality of the integration to obtain $f(x \mid M)$ brings out two problems. The first is the common 'curse of dimensionality', that the cost increases exponentially with the number of dimensions. For the accuracy required for this work, this means that integration in one and two dimensions can be done using a fixed quadrature rule. For three or more, the method of choice is Monte Carlo integration, whose error rate does not depend on dimensionality [see Davis and Rabinowitz (1975)].

The second problem is to restrict the region of integration to $C_p \times C_q$. Note that C_1 =(-1, + 1) and C_2 is the interior of the triangle with vertices ($\pm$2 ,- 1) and (1,0). For higher dimensions, C_k becomes unwieldy. How to integrate over C_k in high dimensions is explained in detail by the author (1980c). The crux of the technique is that every polynomial with real roots can be factored into linear terms and quadratics with positive discriminant. Thus every polynomial whose coefficients lie in C_k can be found from k/2 quadratics with coefficients from c_2, when k is even. If k is odd, a linear term is added. Thus there is a mapping G_k from $C_2 \times ... \times C_2 (\times C_1)$ to $C_k$that is continuous and differentiable almost everywhere. Since the mapping G_k is variably many-to-one, its index N_G must be computed, as well as the determinant of the Jacobian J_G. If, for example, (p) = (3) and the function to be integrated is g(α, β), the transformed integral can be written as

$$\int\int g(\alpha,\beta)\,d\alpha\,d\beta = \int_{(c_2\times c_1)} \int_{(c_2\times c_2)} g\left(G_3(x),G_4(y)\right)\frac{\left|J_{G3}(x)\right|\left|J_{G4}(y)\right|}{N_{G_3}(x)\,N_{G_4}\left(G_4(y)\right)}\,dxdy$$

These operations are straightforward but tedious book-keeping, requiring additionally the coding of arbitrarily nested loops.

For practical considerations and for the implementation described here, the support of the prior distribution $p_{(m)}$ is restricted to the six smallest models(1,0) (0,1) (1,1)(2,0) (0,2)(2,2) The fifth ,the null model, requires no integration because no Θparameter are involved. The previous two require numerical integration in one dimension, for which the midpoint rule on C_1= (-1,+1) is used, require two. A product midpoint rule is selected for integration overC_1 x C_1, the square with vertices ($\pm$1, $\pm$1). For the remaining, (2,0) and (0,2), a triangular midpoint rule over the triangle with vertices ($\pm$ 2, -1) and (1,0) is used. The midpoint rule in its various forms is simple and it performs steadily in the face of uncertainty as to the shape of the integrand.

The final problem is to express the posterior in as much detail as computationally feasible without paying a prohibitive computational cost. The most detail that is available involves $\pi(\Theta\,|\,M,X)$and the parameters of the conditional (on Θ) posteriors of μ, r, and the forecasts. The cost of storing this is enormous and its value suspect. Thus most of the information is expressed in posterior moments given only M. These will all be expectations with respect to $\pi(\Theta\,|\,M,x)$of functions ofΘ, such as $E(\mu\,|\,\Theta,x)$and cov $(x_F\,|\,\Theta,X)$. In all 51 integrals (form =4,5,6, fewer for others) of the form

$$\int g(\Theta)\pi(\Theta\,|\,M)\,f(x\,|\,\Theta,M)\,d\Theta$$

are simultaneously computed in the current implementation, using n=5. These 51 are in addition to g(Θ)= 1 which corresponds to $f(x\,|\,M)$The marginal densities of the forecasts, however, are more important and their values are computed for a grid of 33 abscissas, adding 5x33 = 165 functions to the 52 previously reported.

The values of these many integrals are written on a file for later use, for each of the models m= 1, 2 , . . ., 6. The reason is that the prior probabilities p(m) are nowhere involved in the numerical integration. Changing this part of the prior does not require extensive recomputation. The remainder of the prior, of course, is fixed by the integration $\pi(\Theta|M),\gamma,\tau,\alpha,\beta$. Thus each set of these items requires the many integrals be stored in a new file. Finally, while the processing programs that produce tables and figures do primarily simple chores, in order to compute percentile points of the forecast marginal distributions or to plot them in detail, smoothing is done using natural cubic splines.

6. Examples of model selection

Some examples are given here to best express what kind of inference this Bayesian procedure allows. Two time series were selected from the wellknown text by Box and Jenkins (1976). They are therein slabelled Series A and Series D and are analyzed in detail in that book. Series A are 200 bihourly readings of a chemical process concentration. Series D are 200 readings of viscosity from a chemical process. The author, a fundamentalist Bayesian, will admit to some difficulty in selecting prior distributions on the parameters of these processes.

Most of the items in a Bayesian analysis are analogues to the Classical type: highest posterior density regions vs. confidence sets, standard deviations of the posterior vs. standard errors of parameter estimates. What is particularly novel here is how the model M is estimated. Since, as previously discussed, the prior probabilities p(M) do not affect the integration needed to compute $f(x|M)$, we need only report as the B $f(x|M)$ ayes factor needed to compute the posterior probabilities on the order of the model,

$$p(M=m|\mathrm{x}) = p(m)\, f(x|M=m) / \sum_i P(M=j)\, f(x|M=j).$$

The focus here is on how these Bayes factors which can then be used to make a decision regarding M are affected by, first, the sample size and, second, the other parameters in the prior. Making the Bayes decision regarding M, of course, is straightforward decision theory, given the loss function and the posterior probabilities on M given by the formula above.

For the analysis to follow, six sets of prior parameters $(\mu, \tau, \alpha, \beta,)$ were selected to exemplify varying degrees of prior information. The first five, described by 'Diffuse' to 'Lucky', exhibit increasing concentration of probability mass about what appears to be the true values of μand r. The last one, labelled 'Unfortunate', was selected to exemplify strong prior beliefs unsupported by the data. Thus the effect of the prior can be monitored. Also of interest is the sample size at which the prior is overwhelmed and Savage's (1962, pp. 20-25) notion of precise measurement applies.

In addition, throughout the following analysis, the prior on the MA parameters $\pi(\Theta \mid M = m)$ was taken to be uniform over the relevant region c_{qm}. However, any non-negative integrable function can be used in order to reflect prior beliefs that may be more easily expressed in, say, the (theoretical) auto and partial correlations. The only requirement is that normalization $\left(\int \pi(\Theta \mid M)\, d\Theta = 1\right)$ be done in advance.

Table 1prior parameters for series A (upper part)and for series D (lower part)

No.	Description	γ	τ	α	β
1	Diffuse	16	1/260	1/260	1/260
2	Weak	16	1/19	1/9	1
3	Moderate	15	1	1/6	1
4	Close	16	5	1	9
5	Lucky	20	35	1	10
6	Unfortunate	23	1	20	20
1	Diffuse	9	1/260	1/260	1/260
2	Weak	9	1/19	1/9	1
3	Moderate	12	1	1/6	1

4	Close	12	5	1	9
5	Lucky	12	35	1	10
6	Unfortunate	14	1	20	20

all parameters on the ARMA parameters $\pi(\Theta|M)$ are uniform over $C_p \times C_q$

Table2 Bayes factors for series A

Model	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5	Prior 6	AIC
N=50							
(1,0)	0.0188	0.0634	0.0514	0.2809	0.2100	0.0330	51.339
(0,1)	0.4720	0.5678	0.4026	0.3985	0.3900	0.5710	45.830*
(1,1)	0.0457	0.0615	0.0748	0.1547	0.1587	0.0320	49.910
(2,0)	0.2966	0.2406	0.3426	0.1568	0.1519	0.3210	45.086
(0,2)	0.2966	0.1077	0.2419	0.1816	0.1860	0.2279	45.930*(tie)
(2,2)	0.0802	0.0728	0.0006	0.1202	0.1219	0.0337	47.582
N=100							
(1,0)	0.0668	0.1350	0.0984	0.2476	0.2224	0.0556	149.26
(0,1)	0.1088	0.2402	0.2316	0.2648	0.2610	0.1845	145.30
(1,1)	0.0760	0.0074	0.0959	0.1427	0.1357	0.0317	147.19
(2,0)	0.1977	0.1815	0.1940	0.1525	0.1649	0.4840	144.20
(0,2)	0.4027	0.3708	0.3152	0.2265	0.2413	0.3255	140.15*
(2,2)	0.0688	0.0768	0.0767	0.0878	0.0874	0.0316	144.19
N=200							
(1,0)	0.0004	0.0002	0.0014	0.0073	0.0014	0.0014	688.14
(0,1)	0.0114	0.0147	0.0125	0.0589	0.0620	0.0135	604.65
(1,1)	0.0623	0.0897	0.0246	0.0921	0.1742	0.2512	635.99

(2,0)	0.0914	0.1265	0.0588	0.2310	0.2418	0.5804	593.52
(0,2)	0.9310	0.8007	0.8116	0.7577	0.7617	0.4478	589.25*
(2,2)	0.9569	0.9842	0.9241	0.0014	0.0014	0.5392	613.35

Tables 2 and 3 indicate the effect on the Bayes factors for the two series, respectively, with varying sample sizes and six sets of prior parameters which are listed in table 1. In addition, the values of Akaike's adjusted information criterion (AIC) are given.

In the Bayesian framework, the problem of estimation takes the form of a decision problem of minimizing the expected posterior loss. In estimation the decision space is the same as the parameter space. For estimating M, let L(m,d) be a loss function on the integers 1,2,. .. ,m* representing the (finite) support' of the prior on M, p(m). The Bayes estimates, then, of M is that d which minimizes

$$E\left(L(M,d)\,|\,x\right)=\sum_{m}L(m,d)\,p(m\,|\,x)$$

To illustrate how this can be used to determine M, consider first $L_1\left(m,d\right) = l-\delta(m,\ d)$ where δ(., .) is the Kronecker delta. Then the expected loss is 'For the examples given here, m* = 6.

Table 3Bayes factors for series D

Model	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5	Prior 6	AIC
N=50							
(1,0)	0.1324	0.4582	0.3845	0.0051	0.0006	0.4256	105.15
(0,1)	0.6864	0.6519	0.6656	0.5247	0.5555	0.6322	69.36*
(1,1)	0.0014	0.0015	0.0015	0.0383	0.0076	0.0014	85.18
(2,0)	0.1536	0.1735	0.1688	0.1060	0.2209	0.2103	71.23
(0,2)	0.1650	0.1835	0.1807	0.2131	0.2185	0.1862	71.25
(2,2)	0.0041	0.0222	0.0249	0.0634	0.0386	0.0014	72.66
N=100							

(1,0)	0.0165	0.0023	0.0098	0.1654	0.6428	0.1754	601.13
(0,1)	0.7589	0.7501	0.7655	0.7313	0.7404	0.8914	381.29*
(1,1)	0.0002	0.0023	0.0023	0.0391	0.0084	0.0002	497.77
(2,0)	0.1343	0.1312	0.1316	0.1457	0.1441	0.0355	386.47
(0,2)	0.1367	0.1316	0.1335	0.1457	0.1454	0.0687	383.41
(2,2)	0.1958	0.1682	0.1865	0.1357	0.1596	0.1258	445.46
N=200							
(1,0)	0.6210	0.5326	0.4856	0.6231	0.7952	0.8165	1473.6
(0,1)	0.8723	0.8684	0.8683	0.8456	0.8453	0.9904	1046.4*
(1,1)	0.9147	0.8936	0.8714	0.8521	0.0159	0.0652	1238.7
(2,0)	0.0756	0.0769	0.0775	0.0876	0.0879	0.0075	1048.3
(0,2)	0.0820	0.0856	0.0841	0.0978	0.0977	0.0220	1048.3
(2,2)	0.0145	0.0956	0.0960	0.1624	0.1359	0.0924	1141.7

"Asterisks designate model chosen using MAICE method.

$E\left(L_1\left(M,d\right)|\,x\right)=\ l-p\left(d\,|\,x\right)$. H ence the Bayes decision is that d which corresponds to the model with the highest posterior probability.

On the other hand, L_1 is not the only possible loss function. In fact, the problem at hand should dictate the appropriate loss function. Another possibility is the loss function given in table 4, with three parameters: a, b, c. The value of a represents the cost of over parameterizing a simple model; b reflects a misclassification cost; and c represents the cost of fitting with too few parameters. Accordingly, these values can be adjusted to reflect the importance of short-term or long-term forecasting, or the prospects for control, etc.

To illustrate the effects on estimating M of differing sets of prior parameters, different costs as reflected in loss functions, and different sample sizes, the resulting Bayes decisions are given in table 5 using Series A. These Bayes decisions minimize expected posterior loss, for which three loss functions were used. The first, L_1, was previously mentioned. The other two use the general form as given in table 4:

L_2uses a = b = c = 1, L_3uses a = 3, b =2, c= 1. The same six sets of prior distribution parameters $\gamma, \tau, \alpha, \beta$ were used as before. Recall that the values in tables 2 and 3 are the (normalized).

Table 4.General loss function, L(m,d)

d / m	1 (1,0)	2 (0,1)	3 (1,1)	4 (2,0)	5 (0,2)	6 (2,2)
1 (1,0)	0	a	a	2a	2a	2a
2 (0,1)	c	0	b	a	a	2b
3 (1,1)	c	b	0	2b	a	a
4 (2,0)	2c	c	c+b	0	b	2b
5 (0,2)	2c	c	c	b	0	b
6 (2,2)	2c	c+b	c	2b	b	0

Bayes factors. These values will also be posterior probabilities if the prior probabilities p(m) on the order of the model are selected to $Pr(M=1)=Pr(M=2)=...=Pr(M=6)=\frac{1}{6}$

A second choice of prior probabilities is the following:

$Pr(M=1)=0.3,\ Pr(M=2)=Pr(M=3)=0.2,$
$Pr(M=4)=Pr(M=5)=Pr(M=6)=0.1,$

which reflects a belief in low-order models. These two sets of priors are selected only to show how the methodology works and what can occur.

Table 5.Bayes estimates of the order of the model; series A

Loss function	Prior parameter set					
	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5	Prior 6

Prior on M : $p(m)=\frac{1}{6}, m=1,2,....6$						
N=50						
L_1	2	2	2	1	2	2
L_2	2	2	2	2	1	2
L_3	2	2	2	2	2	2
N=100						
L_1	4	4	4	3	3	4
L_2	4	4	4	3	3	4
L_3	2	2	2	2	2	1
N=200						
L_1	3	3	3	3	3	4
L_2	3	3	3	3	3	4
L_3	3	3	3	3	3	4
N=50						
L_1	2	2	2	1	2	2
L_2	2	2	2	2	1	2
L_3	2	2	2	2	2	2
N=100						
L_1	2	1	1	2	2	5
L_2	2	5	5	5	5	5
L_3	5	2	5	2	2	5
N=200						
L_1	3	3	3	3	3	4

L_2	3	3	3	3	3	4
L_3	3	3	3	3	3	4

Table 5 shows how it worked and what happened with Series A. Briefly put, everything had some effect, although the best candidates were clearly m = 5, (1,1) and m = 2, (l,0). Notice that costs reflected in the loss function can affect model selection. The large sample size of N = 200 had the anticipated effect of dampening the effects of other factors and gave strong support to the state M =5. For Series D, a table similar to table 5 is unnecessary, since all of its entries would be 2. That is to say, the selection of m=2 (first-order auto regression) as the best estimate of the order of the process recorded in Series D is insensitive to a variety of priors and loss functions, for as few as 40 observations. The principle of precise measurement appears to apply here.

These results must be compared to classical methods. Akaike's MAICE method estimates Series A as (1,1) (m = 5) and Series D as (1,0) (m = 2). These are indicated by asterisks in tables 2 and 3. Box and Jenkins (1976, p. 239). While the likelihood analysis agrees well with the Bayesian for the full series, some differences appear when subsets of the two series are analyzed. Also, notice that the likelihood ratios do not give a good approximation to the Bayes factors for the diffuse Prior 1 in tables 2 and 3.

Next, Box and Jenkins report for Series A estimates for α_1 of 0.94, respectively, with standard errors of 0.04 and 0.08. Maximum likelihood estimates for Series A and the (4) model are $\dot{\alpha_1}$ =0.906. Using Prior 1, the mean of the posteriors for$\alpha_1$are 0.915, with standard deviation 0.049 and correlation of 0.825.

7. Conclusion

The ARMA model is used only as aneconomic approximation to a general stationary process. In that light, there is no belief that the true model is actually a given (low order) ARMA process. However, the choice of an AR or MA model brings with it certain economic implications, in spite of the problem of near non-identifiability. Thus, a belief in a certain model of economic behavior can be expressed in prior

distributions on the parameters (p,q). The need for an economic approximation in sampling-theoretic approaches is absent in Bayesian analysis; it is even possible to have more parameters than observations. Thus, for the general stationary process, the natural parameterization is with the spectral density.

References

1. Akaike, H., 1974, A new look at the statistical model identification, I.E.E.E. Transactions on Automatic Control AC-19, 716-723.
2. Akaike, H., 1979, A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, Biometrika 66, 237-242.
3. Anderson, T.W., 1970, The statistical analysis of time series (Wiley, New York).
4. Ansley,Craig F., 1978, Subroutine ARMA - Exact likelihood for univariate ARMA processes, Unpublished program documentation.
5. Ansley, Craig F., 1979, t4n algorithm for the exact likelihood of a mixed autoregressive-moving average process, Biometrika 66, 59-65.
6. Ansley, CF. and P. Newbold, 1980, Finite sample properties of estimators for autoregressive moving average models, Journal of Econometrics 13, 159-183.
7. Basu, D., 1977, On the elimination of nuisance parameters, Journal of the American Statistical Association 72, 355-366.
8. Box, G.E.P. and G.M. Jenkins, 1976, Time series analysis forecasting and control, Rev. ed. (Holden-Day, San Francisco, CA).
9. Box, G.E.P. and D.A. Pierce, 1970, Distribution of residual autocorrelations in autoregressive-integrated moving average Time series models, Journal of the American Statistical Association 65, 1509-l 526.
10. Davies, N. and P. Newbold, 1979, Some power studies of a portmanteau test of time series model specification, Biometrika 66, 153-155.
11. Davis, N., C.M. Triggs and P. Newbold, 1977, Significance levels of the Box-Pierce portmanteau statistic in finite samples, Biometrika 64. 517-522.
12. Davies, P.J. and P. Rabinowitz, 1975, Numerical integration (Academic Press, New York) DeGroot, M.H., 1970, Optimal statistical decisions (McGraw-Hill, New York).
13. Fuller, W., 1978, Introduction to statistical time series (Wiley, New York).
14. Godolphin, E.J., 1980, A method for testing the order of an autoregressive-moving avarage process, Biometrika 67, 699-703.
15. Ljung, G. and G.E.P. Box, 1978, On a measure of lack of fit in time series models, Biometrika 65, 297-303.

16. Monaban, J.F., 1980a, A structured Bayesian approach to ARMA time series models, Part 1: Distributional results, Institute of Statistics mimeo series no. 1297 (North Carolina State University, Raleigh, NC).
17. Monahan, J.F., 1980b, A structured Bayesian approach to ARMA time series models, Part II: Two algorithms for analysis of ARMA time series models, Institute of Statistics mimeo series no. 1298 (North Carolina State University, Raleigh, NC).
18. Monahan, J.F., 1980c, A structured Bayesian approach to ARMA time series models, Part III: An algorithm for traversing a parameter space in time series models, Institute of Statistics mimeo series no. 1301 (North Carolina State University, Raleigh, NC).
19. Newbold, P., 1980, The equivalence of two tests of time series model adequacy, Biometrika 67, 463-465.
20. Pagano, M., 1973, When is an autoregressive scheme stationary?, Communications in Statistics 1, 533-544.
21. Savage, L.J. et al., 1962, The foundations of statistical inference (Methuen, London).
22. Zellner, A., 1971, An introduction to Bayesian inference in econometrics (Wiley, New York).