Minimum Message Length Autoregressive Model Order Selection

Leigh J. Fitzgibbon
School of Computer Science and Software Engineering,
Monash University
Clayton, Victoria 3800, Australia
leighf@csse.monash.edu.au

David L. Dowe
School of Computer Science and Software Engineering,
Monash University
Clayton, Victoria 3800, Australia

Farshid Vahid
Department of Econometrics and Business Statistics,
Monash University
Clayton, Victoria 3800, Australia
Farshid.Vahid@BusEco.monash.edu.au

Abstract
We derive a Minimum Message Length (MML) estimator for stationary and nonstationary autoregressive models using the Wallace and Freeman (1987) approximation. The MML estimator's model selection performance is empirically compared with AIC, AIC_C, BIC and HQ in a Monte Carlo experiment by uniformly sampling from the autoregressive stationarity region. Generally applicable, uniform priors are used on the coefficients, model order and $\log\sigma^2$ for the MML estimator. The experimental results show the MML estimator to have the best overall average mean squared prediction error and best ability to choose the true model order.

Keywords
Minimum Message Length, MML, Bayesian, Information, Time Series, Autoregression, AR, Order Selection.

INTRODUCTION
The Wallace and Freeman (1987) Minimum Message Length estimator (MML87) [22] is an information-theoretic criterion for model selection and point estimation. It has been successfully applied to many problems including (univariate) linear and polynomial regression models in [19, 18, 17, 14] (and sequence data [6, 7], etc. [20, 22, 21]). The purpose of this paper is to investigate the use of the MML87 methodology for autoregressive (time series) models. Autoregressive models differ from standard linear regression models in that they do not regress on independent variables since the regressor is a subset of the dependent variables (i.e., its lagged values) - the independent variable is really time. This has important (philosophical and practical) ramifications for the MML87 regression estimator, which otherwise assumes that the independent variables are transmitted to the receiver up front (or are already known by the receiver). The independent variables appear in the Fisher information matrix, which is necessary for computation of the MML87 coding volume. Such a protocol would be nonsensical if directly applied to autoregression since it corresponds to transmission of the majority of the data, before the data, in order to transmit the data.

In this paper we investigate several MML87 estimators that take these issues into account. In particular, we focus on an MML87 estimator that is based on the conditional likelihood function, using the least squares parameter estimates. This MML estimator's model selection performance is empirically compared with AIC [1], corrected AIC (AIC_C) [10], BIC [15] (or 1978 MDL [13]), and HQ [9] in a Monte Carlo experiment by uniformly sampling from the autoregressive stationarity region. While MML is geared towards inference rather than prediction, we find that choosing the autoregressive model order having the minimum message length gives a prediction error that is superior to the other model selection criteria in the Monte Carlo experiments conducted.

BACKGROUND

Autoregression
Regression of a time series on itself, known as autoregression, is a fundamental building block of time series analysis. A linear autoregression with unconditional mean of zero[1] relates the expected value of the time series linearly to the p previous values:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + e_t, \quad e_t \sim N(0, \sigma^2) \quad (1)$$

An order p autoregression (AR(p)) model has $p+1$ parameters: $\theta = (\phi_1, ..., \phi_p, \sigma^2)$. Some example data generated from various AR(2) models are displayed in Figures 2 to 7. The examples chosen are all stationary with all inside, but many close to, the boundary of the stationarity region (see Figure 1) to illustrate some of the diversity that can be found in an autoregression model.

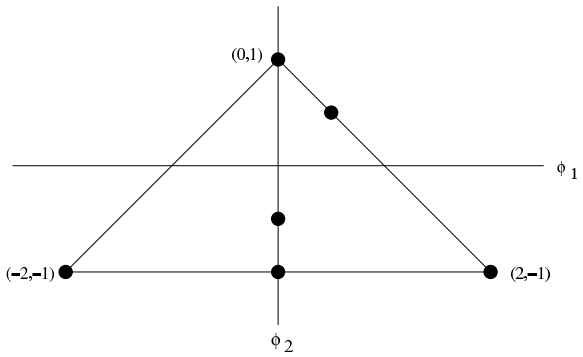

Figure 1. AR(2) stationarity region with the plotted example points (see Figures 2 to 7 on next page) identified.

[1] In this paper we only consider zero mean autoregressions for simplicity. This does not cause any loss of generality, and all results would be qualitatively the same without this assumption.

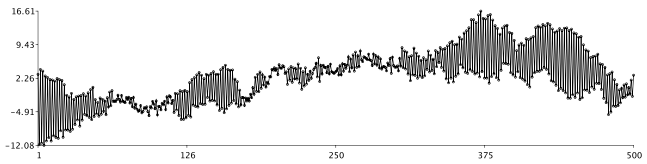

Figure 2. Example data: $\phi_1 = 0$, $\phi_2 = 0.99$ and $\sigma^2 = 1$.

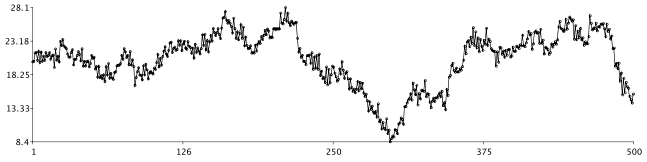

Figure 3. Example data: $\phi_1 = 0.499$, $\phi_2 = 0.499$ and $\sigma^2 = 1$.

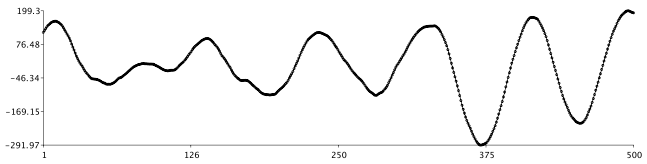

Figure 4. Example data: $\phi_1 = 1.99$, $\phi_2 = -0.995$ and $\sigma^2 = 1$.

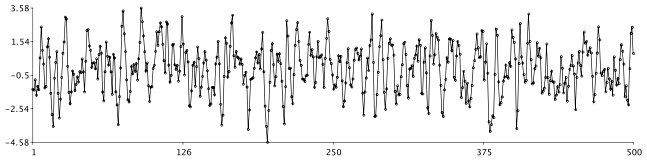

Figure 5. Example data: $\phi_1 = 0$, $\phi_2 = -0.5$ and $\sigma^2 = 1$.

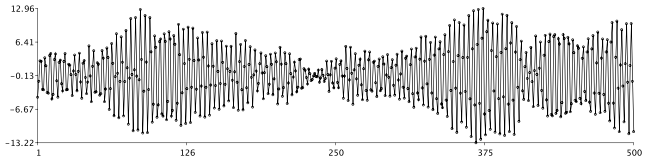

Figure 6. Example data: $\phi_1 = 0$, $\phi_2 = -0.99$ and $\sigma^2 = 1$.

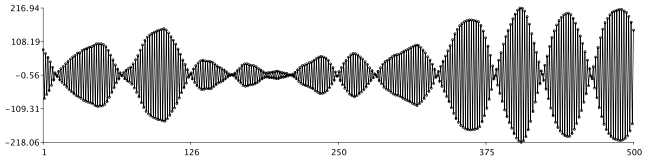

Figure 7. Example data: $\phi_1 = -1.99$, $\phi_2 = -0.995$ and $\sigma^2 = 1$.

The conditional[2] negative log-likelihood of θ is:

$$-\log f(y_{p+1}, ..., y_T|\theta, y_1, .., y_p) = \frac{T-p}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=p+1}^{T}(y_i - \phi_1 y_{i-1} - ... - \phi_p y_{i-p})^2 \quad (2)$$

It is common to make the assumption that y_t is weakly stationary. This assumption means that the time series has a constant mean (in our case zero) and the autocovariances depend only on the distance in time between observations and not the times themselves:

$$E(y_t) = 0 \quad \forall t \quad (3)$$

$$E(y_t y_{t-j}) = E(y_t y_{t+j}) = \gamma_j \quad \forall t \forall j \quad (4)$$

The first p γ's can be calculated as the first p elements of the first column of the p^2 by p^2 matrix $\sigma^2(I_{p^2} - (F \otimes F))^{-1}$ [8, page 59], where F is defined as [8]:

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad (5)$$

The stationarity assumption implies that the first p values are distributed as a multivariate Normal distribution [8, page 124]. The negative log-likelihood of θ for the first p values, $z_p = [y_1, y_2, \dots, y_p]'$, is:

$$-\log f(y_1, ..., y_p|\theta) = \frac{p}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|V_p^{-1}| + \frac{1}{2\sigma^2} z_p' V_p^{-1} z_p \quad (6)$$

where $\sigma^2 V_p$ is the p by p autocovariance matrix.
The exact negative log-likelihood function, when stationarity is assumed, is therefore (using equations 6 and 2):

$$-\log f(y|\theta) = -\log f(y_1, ..., y_p|\theta) - \log f(y_{p+1}, ..., y_T|\theta, y_1, .., y_p) \quad (7)$$

Parameter Estimation

There are a large number of methods used to estimate the parameters of an autoregressive model (see, e.g., [8, 3, 4]). Some of these methods include:

- Least squares
- Conditional least squares (OLS)
- Yule-Walker
- Burg
- Maximum likelihood

Maximum likelihood is rarely used because its solution is non-linear, requiring a numerical solution which is slow and can suffer from convergence problems [4]. The estimates provided by the Yule-Walker, Burg and maximum likelihood methods are guaranteed to be stationary whereas the least

[2] conditional on the first p observed values.

squares estimates are not [4]. Conditional least squares estimates are often preferred since they are easily computed and consistent regardless of whether the data is considered to be stationary or not [8, page 123].

Model Order Selection

Some methods that are commonly used to automatically select the order of an autoregression are AIC [1], corrected AIC (AIC_C) [10], BIC [15] (the same as 1978 MDL [13], as noted below), and HQ [9]:

$$\text{AIC}(p) = \log(\hat{\sigma}_p^2) + \frac{2p}{T} \quad (8)$$

$$\text{AIC}_\text{C}(p) = \log(\hat{\sigma}_p^2) + \frac{2(p+1)}{T-p-2} \quad (9)$$

$$\text{BIC}(p) = \text{MDL}(p) = \log(\hat{\sigma}_p^2) + \frac{p\log(T)}{T} \quad (10)$$

$$\text{HQ}(p) = \log(\hat{\sigma}_p^2) + \frac{2p\log\log(T)}{T} \quad (11)$$

where $\hat{\sigma}_p^2$ is the maximum likelihood estimate of the noise variance for the pth order model. In practice, the exact maximum likelihood estimate is rarely used. Instead, one of the other methods mentioned in the previous section is used as an approximation. A robust method that can be used for stationary and nonstationary data [16] is to use the OLS estimate for $\phi_1, ..., \phi_p$ and estimate σ_p^2 using

$$\hat{\sigma}_p^2 = \frac{1}{T-p}\sum_{i=p+1}^{T}\left(y_i - (\hat{\phi}_1 y_{i-1} + ... + \hat{\phi}_p y_{i-p})\right)^2 \quad (12)$$

Minimum Message Length Inference

The Minimum Message Length (MML) principle [20, 22, 21] is a method of inductive inference, and encompasses a large class of approximations and algorithms. In this paper we use the popular MML87 approximation [22], which approximates the message length for a model consisting of several continuous parameters $\theta = (\theta_1, \ldots, \theta_n)$ by:

$$\text{MessLen}(\theta, y) = -\log\left(\frac{h(\theta)f(y_1, ..., y_N|\theta)\epsilon^N}{\sqrt{|I(\theta)|}}\right) + \frac{n}{2}(1 + \log\kappa_n) - \log h(n) \quad (13)$$

where $h(\theta)$ is a prior distribution over the n parameter values, $f(y_1, ..., y_N|\theta)$ is the standard statistical likelihood function, $I(\theta)$ is the expected Fisher Information matrix, κ_n is a lattice constant ($\kappa_1 = 1/12$, $\kappa_2 = 5/(36\sqrt{3})$, $\kappa_3 = 19/(192 \times 2^{1/3})$, $\kappa_n \to 1/(2\pi e)$ as $n \to \infty$) [5, page 61] [22, sec. 5.3] which accounts for the expected error in the log-likelihood function due to quantisation of the n-dimensional space, ϵ is the measurement accuracy of the data, and $h(n)$ is the prior on the number of parameters. The MML87 message length equation, Equation 13, is used for model selection and parameter estimation by choosing the model order and parameters that minimise the message length. (For a comparison between MML and the subsequent Minimum Description Length (MDL) principle [13], see [21]).

MML87 AUTOREGRESSION

In this section we derive several MML87 message length expressions for an autoregression model. The first decision that must be made is the form of the MML message. There are two seemingly obvious message formats that we could use based on the conditional or exact likelihood functions. If the exact likelihood function is used we are assuming that the data comes from a stationary process. If the conditional likelihood function (Equation 2) is used, stationarity need not be assumed, but we must transmit the first p elements ($\{y_1, ..., y_p\}$) of the series prior to transmission of the data. In this section we give three message formats. The first uses the exact likelihood function and thus assumes stationarity, the second uses the conditional likelihood function and does not assume stationarity, and the third is a combination of the previous two.

Stationary Message Format

The format of the MML87 message using the exact likelihood function (Equation 7) is as follows:

1. First part: $p, \phi_1, ..., \phi_p, \sigma^2$
2. Second part: $y_1, ..., y_T$

MML87 requires the determinant of the expected Fisher information matrix in order to determine the uncertainty volume for the continuous parameters. We can write the expected Fisher information as the sum of two terms which arise in the exact likelihood function (Equation 7 from Equations 6 and 2):

$$I(\theta) = I_{y_1,...,y_p}(\theta) + I_{y_{p+1},...,y_T}(\theta) \quad (14)$$

The expected Fisher information is easily calculated for the conditional likelihood term:

$$I_{y_{p+1},...,y_T}(\theta) = \begin{bmatrix} \sigma^{-2}E(X'X) & 0 \\ 0 & \frac{T-p}{2\sigma^4} \end{bmatrix} \quad (15)$$

where X is a $(T-p)$ by p matrix:

$$X = \begin{bmatrix} y_p & y_{p-1} & \cdots & y_1 \\ y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ y_{T-1} & y_{T-2} & \cdots & y_{T-p} \end{bmatrix} \quad (16)$$

Since we have assumed stationarity, $E(y_t y_{t-j}) = \gamma_j$, and therefore:

$$I_{y_{p+1},...,y_T}(\theta) = \begin{bmatrix} (T-p)V_p & 0 \\ 0 & \frac{T-p}{2\sigma^4} \end{bmatrix} \quad (17)$$

The Fisher information for the multivariate Normal term (recalling Equation 6), $I_{y_1,...,y_p}(\theta)$, is difficult to calculate and $I(\theta)$ can be approximated as $\frac{T}{T-p}I_{y_{p+1},...,y_T}(\theta)$ (see, e.g., [3, page 303]). Using this approximation, the determinant of

the expected Fisher information is equal to:

$$|I(\theta)| \approx \frac{T^{p+1}}{2\sigma^4}|V_p| \quad (18)$$

The stationary model message length is calculated by substitution of Equations 7 and 18 into Equation 13 along with the selection of a suitable prior (we give a generally applicable prior in a later section).

Nonstationary Message Format

In this subsection we give the format of the MML message based on the conditional likelihood function (Equation 2). When the conditional likelihood function is used we must transmit the first p elements, $\{y_1, ..., y_p\}$, of the series to the receiver prior to transmission of (the rest of) the data. The format of the message is therefore:

1. First part: $p, y_1, ..., y_p, \phi_1, ..., \phi_p, \sigma^2$
2. Second part: $y_{p+1}, ..., y_T$

We see that the initial values are transmitted before the continuous parameters, therefore they can then be used to determine the optimal uncertainty volume for the continuous parameters. However, the downside of this ordering is that the receiver must be able to decode the first p values of the time series in order to be able to decode both the parameters and (remaining) data. The initial values must therefore be encoded without such knowledge. We assume that the receiver knows the minimum ($ymin$) and maximum ($ymax$) of the data[3]. We can then encode the initial p values independently, using a uniform density over the interval $[ymin - \epsilon/2, ymax + \epsilon/2]$ where ϵ is the measurement accuracy of the data, in a message of length:

$$\text{MessLen}(y_1, ..., y_p) = p \log\left(1 + \frac{ymax - ymin}{\epsilon}\right) \quad (19)$$

Since the data elements in the message are also encoded to an accuracy of ϵ (recall Equation 13) we find the term $T \log \epsilon$ appearing in the overall message length and therefore (for sufficiently small ϵ) the choice of ϵ does not significantly affect model order selection.

Calculation of the expected Fisher information for the conditional likelihood function in the nonstationary case is difficult. Instead, we use the partial expected Fisher:

$$I_{y_{p+1},...,y_T}(\theta) \approx \begin{bmatrix} \sigma^{-2}X'X & 0 \\ 0 & \frac{T-p}{2\sigma^4} \end{bmatrix} \quad (20)$$

The nonstationary model message length is then the sum of Equation 19 and Equation 13 after substitution of Equations 2 and 20 and the selection of a suitable prior (we give a generally applicable prior in a later section).

Combining the Two Formats (Combination Format)

A slightly more efficient message can be constructed by using a combination of the stationary and nonstationary messages. The sender can compute the length of the transmission using both messages and then choose to use the one having the shorter message length. An additional bit is required at the beginning of the message to tell the receiver which message format has been used. In practice we will generally be using the OLS estimate since the proper MML estimate requires a non-linear solution. For the OLS estimate this coding scheme can be translated to the following procedure:

- Compute the OLS estimate.
- Determine if the estimated model is stationary (e.g. by checking the eigenvalues of F, Equation 5).
- If nonstationary then the nonstationary message must be used because the assumptions made in the multivariate Normal term (Equation 6) of the likelihood function are no longer valid
- Else (if stationary) use the shorter of the two messages.

BAYESIAN PRIORS

Supposing that we are ignorant prior to observation of the data, but expect the data to be stationary, we place a uniform[4] prior on $p, \phi_1, ..., \phi_p$ and $\log \sigma^2$:

$$h(p) \propto 1 \quad (21)$$

$$h(\phi_1, ..., \phi_p, \sigma^2) \propto \frac{1}{R_p} \times \frac{1}{\sigma^2} = \frac{1}{R_p \sigma^2} \quad (22)$$

where R_p is the hypervolume of the stationarity region. Based on results in [2], [12] gives the following recursion for computation of the stationarity region hypervolume:

$$\begin{aligned} M_1 &= 2 \\ M_{p+1} &= \frac{p}{p+1} M_{p-1} \\ R_p &= (M_1 M_3 \times ... \times M_{p-1})^2 \quad \text{for } p \text{ even} \\ R_{p+1} &= R_p M_{p+1} \end{aligned}$$

The stationarity region hypervolume is plotted in Figure 8 for orders 1 to 30.

EMPIRICAL COMPARISON

In the following experiments we have used the conditional ordinary least squares (OLS) estimates and concentrated on the nonstationary message format (i.e., the second format)[5] using the priors given in the previous section. The nonstationary MML format is labelled 'MML' in the graphs and tables. We have used the OLS estimates because they come closest to the true minimum message length estimates. When data is generated we simulate $1000 + T$ elements and then dispose of the initial 1000 elements. The residual variance estimate, $\hat{\sigma}^2$, is calculated using Equation 12 (where T is

[3] or equivalently that these values have been transmitted using a general code, thus adding a constant to the message length.

[4] We note that the ranges used on p and $\log \sigma^2$ do not affect the order selection experiments conducted in this paper but would need to be specified in some modelling situations.

[5] While the combination message format was found to provide a slight improvement over the nonstationary message format, it did not fit into our experimental protocol (i.e., it could be considered to have an unfair advantage over the other criteria by gaining information from the multivariate Normal terms appearing in the exact likelihood function). It is also less practical since it requires significantly more computation time.

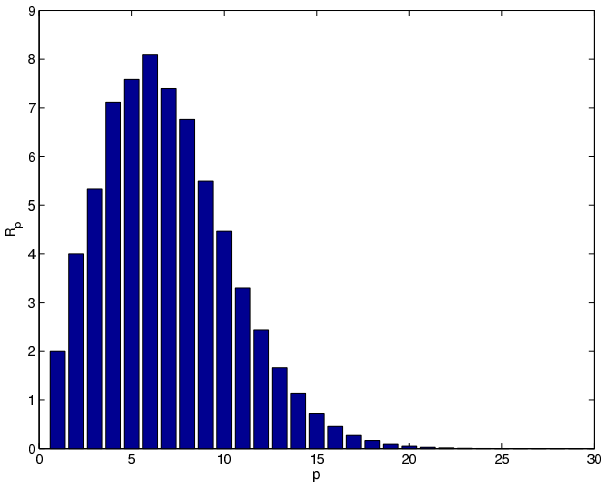

Figure 8. Plot of the hypervolume of the stationarity region for orders 1 to 30.

the total sample size). The mean squared prediction error (MSPE) is used to measure performance:

$$\text{MSPE}(p) = \frac{1}{T}\sum_{i=T+1}^{2T}\left(y_i - (\hat{\phi}_1 y_{i-1} + ... + \hat{\phi}_p y_{i-p})\right)^2 \tag{23}$$

AR(1) Experiments

Data were generated from an AR(1) model for varied values of ϕ_1 over the stationarity region, with $T = 50$. Each method was required to select the model order from the set of OLS estimates for orders 0 to 20. The results can be seen in Figure 9. Each point represents the average MSPE for 1000 data-sets. We see that MML has the smallest average MSPE, especially when the signal is strong (i.e., for $|\phi| \approx 1$).

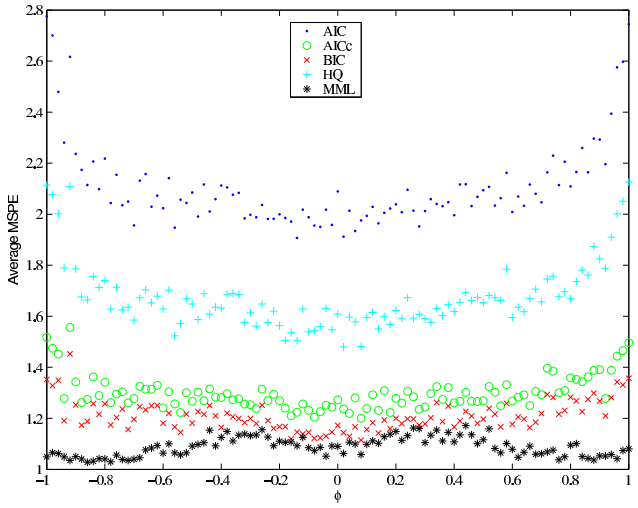

Figure 9. Average Mean Squared Prediction Error for 1000 data-sets for varied true $\phi = \phi_1$ with $T = 50$.

AR(p) Experiments

We conducted experiments for $p = 0$ to $p = 10$ where 1000 autoregression models were sampled uniformly from the stationarity region (see [11] for a means of sampling). Each method was required to select the model order from the set of OLS estimates for orders 0 to p_{max}. The results for $p_{max} = 12$ and $T = 30$ can be found in Table 1 and Figure 10. Results for $p_{max} = 20$ and $T = 50$ can be found in Table 2 and Figure 11. The tables include both frequency counts indicating the number of times each method under/correctly/over inferred the model order and the standard deviation of the MSPE. Each figure and table clearly shows that the MML criterion has the best overall MSPE and best ability to choose the correct model order.

Table 1. AR(p) Simulation result totals for $T = 30$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Average MSPE
AIC	1235	1174	8591	2.8070 ± 5.5384
AICc	3899	2893	4208	2.0509 ± 4.7500
BIC	3307	2515	5178	2.3316 ± 5.2080
HQ	1852	1652	7496	2.6505 ± 5.4838
MML	5999	3196	1805	1.7101 ± 1.6681

Table 2. AR(p) Simulation result totals for $T = 50$

Criterion	$\hat{p} < p$	$\hat{p} = p$	$\hat{p} > p$	Average MSPE
AIC	862	1113	9025	2.5059 ± 2.7529
AICc	3006	3445	4549	1.6007 ± 1.5850
BIC	4150	3821	3029	1.6251 ± 1.6409
HQ	1992	2385	6623	2.1655 ± 2.8760
MML	5713	4314	973	1.2956 ± 0.6682

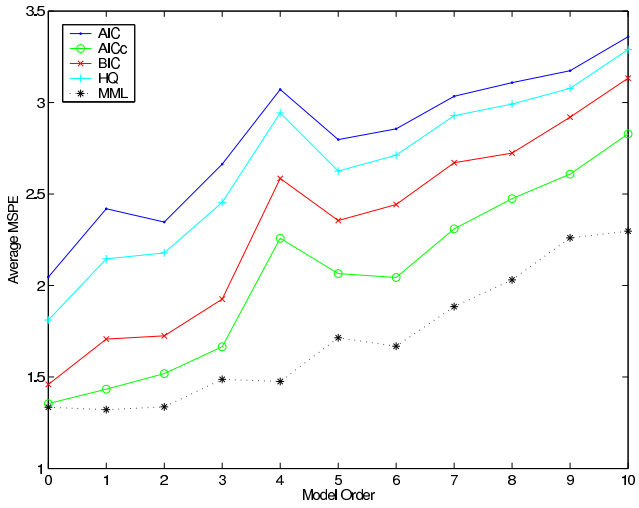

Figure 10. Average Mean Squared Prediction Error for 1000 data-sets for varied true p with $T = 30$.

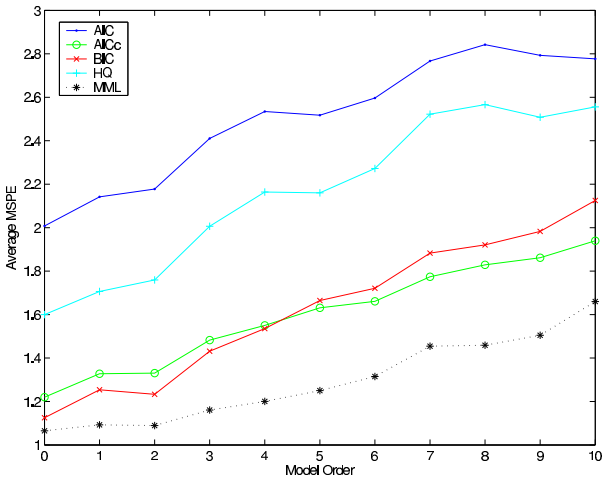

Figure 11. Average Mean Squared Prediction Error for 1000 data-sets for varied true p with $T = 50$.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Australian Research Council (ARC) Discovery Grant DP0343650.

CONCLUSION

We have investigated autoregressive modelling in the MML framework using the Wallace and Freeman (1987) approximation. Three message formats were formulated. The most appropriate format for general use is the nonstationary format, which is based on the conditional likelihood function. This format is simple, easily computed and can be applied regardless of whether the data is stationary or not. When used in conjunction with least squares estimates the MML estimator was found to have very good performance (in squared error and order selection), as the experimental results show.

REFERENCES

[1] Akaike, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, 6 (1974), 716–723.

[2] Barndorff-Nielsen, O., and Schou, G. On the parametrization of autoregressive models by partial autocorrelations. Journal of Multivariate Analysis 3 (1973), 408–419.

[3] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. Time Series Analysis: Forecasting and Control. Prentice Hall, Englewood Cliffs, New Jersey, 1994.

[4] Broersen, P. M. T., and de Waele, S. Empirical time series analysis and maximum likelihood estimation. In IEEE Benelux Signal Processing Symposium (2000), pp. 1–4.

[5] Conway, J. H., and Sloane, N. J. A. Sphere Packings, Lattices and Groups. Springer-Verlag, New York, 1999.

[6] Edgoose, T., and Allison, L. MML Markov classification of sequential data. Journal of Statistics and Computing 9 (1999), 269–278.

[7] Fitzgibbon, L. J., Allison, L., and Dowe, D. L. Minimum message length grouping of ordered data. In Algorithmic Learning Theory (Berlin, 2000), vol. 1968 of LNAI, Springer-Verlag, pp. 56–70.

[8] Hamilton, J. D. Time Series Analysis. Princeton University Press, New Jersey, 1994.

[9] Hannan, E. J., and Quinn, B. G. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B (Methodological) 41, 2 (1979), 190–195.

[10] Hurvich, C. M., and Tsai, C. Regression and time series model selection in small samples. Biometrika 76 (1989), 297–307.

[11] Jones, M. C. Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models. Journal of Applied Statistics 36, 2 (1987), 134–138.

[12] Piccolo, D. The size of the stationarity and invertibility region of an autoregressive-moving average process. Journal of Time Series Analysis 3, 4 (1982), 245–247.

[13] Rissanen, J. J. Modeling by shortest data description. Automatica 14 (1978), 465–471.

[14] Rumantir, G. W., and Wallace, C. S. Sampling of highly correlated data for polynomial regression and model discovery. In Advances in Intelligent Data Analysis (Mar. 2001), vol. 2189, pp. 370–377.

[15] Schwarz, G. Estimating the dimension of a model. The Annals of Statistics 6 (1978), 461–464.

[16] Tsay, R. S. Order selection in nonstationary autoregressive models. The Annals of Statistics 12, 4 (1984), 1425–1433.

[17] Vahid, F. Partial pooling: A possible answer to "To pool or not to pool". In Cointegration, Causality, and Forecasting: A Festschrift in honour of Clive W. J. Granger, R. F. Engle and H. White, Eds. Oxford University Press, New York, 1999. Chapter 17.

[18] Viswanathan, M., and Wallace, C. S. A note on the comparison of polynomial selection methods. In Workshop on Artificial Intelligence and Statistics (Jan. 1999), Morgan Kauffman, pp. 169–177.

[19] Wallace, C. S. On the selection of the order of a polynomial model. Tech. rep., Royal Holloway College, London, 1997.

[20] Wallace, C. S., and Boulton, D. M. An information measure for classification. Computer Journal 11, 2 (Aug. 1968), 185–194.

[21] Wallace, C. S., and Dowe, D. L. Minimum message length and Kolmogorov complexity. The Computer Journal 42, 4 (1999), 270–283.

[22] Wallace, C. S., and Freeman, P. R. Estimation and inference by compact encoding. Journal of the Royal Statistical Society. Series B (Methodological) 49 (1987), 240–252.