Minimum Message Length Order Selection and Parameter Estimation of Moving Average Models

Daniel F. Schmidt

Centre for MEGA Epidemiology, The University of Melbourne
Carlton VIC 3053, Australia
dschmidt@unimelb.edu.au

Abstract. This paper presents a novel approach to estimating a moving average model of unknown order from an observed time series based on the minimum message length principle (MML). The nature of the exact Fisher information matrix for moving average models leads to problems when used in the standard Wallace–Freeman message length approximation, and this is overcome by utilising the asymptotic form of the information matrix. By exploiting the link between partial autocorrelations and invertible moving average coefficients an efficient procedure for finding the MML moving average coefficient estimates is derived. The MML estimating equations are shown to be free of solutions at the boundary of the invertibility region that result in the troublesome "pile-up" effect in maximum likelihood estimation. Simulations demonstrate the excellent performance of the MML criteria in comparison to standard moving average inference procedures in terms of both parameter estimation and order selection, particularly for small sample sizes.

1 Introduction

Moving average models are one of the fundamental building blocks in linear time series analysis. A time series of length n, $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$, is generated by a q^*-th order moving average model with coefficients, $\boldsymbol{\eta}^*_{q^*} = (\eta^*_1, \ldots, \eta^*_{q^*})'$, if

$$y_t = \sum_{j=1}^{q^*} \eta^*_j v_{t-j} + v_t, \tag{1}$$

where $v_t \sim N(0, \tau^*)$ are the independently and identically distributed normal innovations with variance τ^*. The moving average model describes a time series as being composed of a linear combination of q^* unobserved random variables from the series v_t. In general, only the time series $\mathbf{y}$ is available for observation, and the order and parameters must be estimated on the basis of the data alone. A common approach to this problem is to combine maximum likelihood estimation of the parameters with an information criterion based order selection procedure.

Let $\boldsymbol{\theta}_q = (\boldsymbol{\eta}_q', \tau)'$ denote the full parameter vector of a q-th order moving average. To estimate a moving average model using an information criterion one solves

$$\hat{q} = \underset{q \in \{0,\ldots,Q\}}{\arg\min} \left\{ -\log p_q(\mathbf{y}|\hat{\boldsymbol{\theta}}_q) + \alpha(q,n) \right\}, \tag{2}$$

where $p_q(\mathbf{y}|\boldsymbol{\theta}_q)$ is the likelihood of data $\mathbf{y}$ under a moving average model with parameters $\boldsymbol{\theta}_q$,

$$\hat{\boldsymbol{\theta}}_q = \underset{\boldsymbol{\theta}_q \in \Lambda_q \times \mathbb{R}_+}{\arg\max} \left\{ p_q(\mathbf{y}|\boldsymbol{\theta}_q) \right\}$$

is the maximum likelihood estimator of the parameters, and $\alpha(q,n)$ is a complexity penalty function; common choices include $\alpha(q,n) = q$ (Akaike's information criterion [1]) and $\alpha(q,n) = (q/2)\log n$ (the Bayesian information criterion [2]). The set Λ_q is the invertibility region for a q-th order moving average model, i.e., the set of all coefficients $\boldsymbol{\eta}_q$ for which the roots of the characteristic polynomial $1 + \sum_{j=1}^{q} \eta_j z^{-j}$ lie completely within the unit circle.

This paper examines the problem of estimating moving average models using the information theoretic minimum message length (MML) principle [3]. The MML principle has previously been applied to the problem of order selection of moving average models in [4] by Sak et al. Unfortunately, their derivation of the Fisher information matrix contains a serious mistake, and the resulting message lengths are based on a quantity that is neither the exact nor the asymptotic Fisher information. The main contribution of this paper, which is based in part on unpublished work presented in the author's PhD thesis [5], is the derivation of a correct message length formula for moving average models that is based on the asymptotic Fisher information matrix. The details of this new criterion, and some important properties, are discussed in Section 2, and its performance is compared against several moving average inference procedures in Section 3.

1.1 The Minimum Message Length Principle

The MML principle is based on the intimate connection between statistical inference and data compression and has close links to deep concepts such as Solomonoff's algorithmic information theory [6]. Under the MML principle, the explanation that most concisely describes the data is considered the *a posteriori* most likely; as the compression of the data must be decodable, the details of the model used to compress the data must also be included in the description. The length of the compressed data, usually expressed terms of base-e digits, or *nits*, acts as a universal measure of a model's goodness-of-fit that naturally takes into account both the capability and the complexity of the model. While calculation of the exact message length is in general an NP-hard problem [7], the Wallace–Freeman MML87 approximation [8] offers a tractable alternative under some regularity conditions involving the likelihood function and prior distribution (see pp. 226–227, [3]). Let $\omega \in \Omega$ denote a model class in a set of candidate model classes Ω. The MML87 message length for data $\mathbf{y}$ compressed using a fully specified model $\boldsymbol{\theta}_\omega \in \Theta_\omega$, with k_ω continuous parameters, is given by

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}_\omega, \omega) = -\log p_\omega(\mathbf{y}|\boldsymbol{\theta}_\omega) + \frac{1}{2}\log|\mathbf{J}(\boldsymbol{\theta}_\omega)| - \log \pi(\boldsymbol{\theta}_\omega, \omega) + c(k_\omega), \tag{3}$$

where $\pi(\cdot)$ is a joint prior distribution over the parameter space $\mathit{\Theta}_\omega$ and the set of candidate model classes $\mathit{\Omega}$, $\mathbf{J}(\cdot)$ is the Fisher information matrix, and

$$c(k) = -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1) \tag{4}$$

where $\psi(\cdot)$ is the digamma function. Inference is performed by seeking the pair $(\hat{\boldsymbol{\theta}}^{87}_{\hat{\omega}}, \hat{\omega}^{87})$ that minimises the message length (3); in contrast to the information criterion approach, there is no need to appeal to different principles for parameter estimation and order selection.

2 Message Lengths of Moving Average Models

The ingredients required to evaluate the MML87 message length (3) are the likelihood function, a prior distribution over the continuous and structural parameters, and the Fisher information matrix. Data arising from model (1) can be exactly characterised as being generated by an n-dimensional multivariate normal distribution with zero mean and a special covariance matrix $\tau\boldsymbol{\mathit{\Gamma}}(\boldsymbol{\eta}_q)$, with entries $\mathit{\Gamma}_{i,j}(\boldsymbol{\eta}_q) = \mathbb{E}\left[y_i y_j\right]/\tau = \gamma_{|i-j|}(\boldsymbol{\eta}_q)$, where

$$\gamma_k(\boldsymbol{\eta}_q) = \begin{cases} \displaystyle\sum_{j=0}^{q-k} \eta_j \eta_{j+k} & k \le q \\ 0 & k > q \end{cases}, \tag{5}$$

with $\eta_0 = 1$ [9]. The negative log-likelihood of a time series, $\mathbf{y}$, given a parameter vector $\boldsymbol{\theta}_q = (\boldsymbol{\eta}'_q, \tau)'$, is

$$-\log p_q(\mathbf{y}|\boldsymbol{\theta}_q) = \left(\frac{n}{2}\right)\log(2\pi\tau) + \frac{1}{2}\log|\boldsymbol{\mathit{\Gamma}}(\boldsymbol{\eta}_q)| + \left(\frac{1}{2\tau}\right)\mathbf{y}'\boldsymbol{\mathit{\Gamma}}^{-1}(\boldsymbol{\eta}_q)\mathbf{y}. \tag{6}$$

Direct evaluation of (6) involves $O(n^3)$ operations, and therefore becomes infeasible for large sequences. An alternative, and computational efficient approach, is to evaluate the likelihood using the Kalman filter, which involves only $O(n)$ operations; see, e.g., [10] for details[1].

The prior distribution for the moving average coefficients is taken to be uniform over the invertibility region $\mathit{\Lambda}_q$ (as in [11, 4, 5]), the prior distribution for the innovation variance τ is taken to be scale-invariant over some suitable interval (τ_0, τ_1), and the prior distribution for the model order is taken to be uniform over the set $\{0, \ldots, Q\}$, i.e.,

$$\begin{aligned} \pi(\boldsymbol{\eta}_q, \tau, q) &= \pi(\boldsymbol{\eta}_q)\pi(\tau)\pi(q), \\ \pi(\boldsymbol{\eta}_q) &= \frac{1}{\mathrm{vol}(\mathit{\Lambda}_q)}, \\ \pi(\tau) &\propto \frac{1}{\tau}, \\ \pi(q) &\propto 1. \end{aligned} \tag{7}$$

[1] There is a minor typographical error in the initialisation algorithm in [10] in which the second matrix $\mathbf{T}$ is incorrectly transposed in Equation 12.

As the bounds (τ_0, τ_1), and the maximum order Q, appear only as constants in the final message length expression their values have no effect on order selection or parameter estimation and may be safely ignored. For completeness, the algorithm [12] for computing $\mathrm{vol}(\Lambda_q)$ is presented in Appendix A.

2.1 Fisher Information Matrix

By exploiting the fact that the moving average model is a multivariate Gaussian distribution, the exact, finite sample, Fisher information matrix $\mathbf{J}_n(\boldsymbol{\theta}_q)$ may be found using standard formulae. Unfortunately, there are two problems with the exact information matrix: (i) even using fast algorithms, such as the one in [9], computation of the exact information matrix is slow, requiring $O(n^2)$ operations (except in the special case that $q = 1$ [13]), and (ii) the exact information matrix is singular at the boundaries of the invertibility region. The latter problem arises from identifiability issues in the moving average model and can lead to serious violations of the regularity conditions under which the MML87 approximation was derived. Instead, we consider the asymptotic information matrix

$$\mathbf{J}(\boldsymbol{\theta}_q) = n \cdot \lim_{n \to \infty} \left\{ \frac{\mathbf{J}_n(\boldsymbol{\theta}_q)}{n} \right\}.$$

The entries of the asymptotic information matrix are given by Whittle's asymptotic formula [14], and in the case of moving average models they are straightforward to calculate. Define the auxilliary autoregressive process

$$x_t + \sum_{j=1}^{q} \eta_j x_{t-j} = u_t, \tag{8}$$

where $u_t \sim N(0,1)$ are independently and identically distributed normal innovations. The asymptotic information matrix is then given by

$$\mathbf{J}(\boldsymbol{\theta}_q) = n \cdot \begin{pmatrix} \boldsymbol{\Phi}(\boldsymbol{\eta}_q) & 0 \\ 0 & \dfrac{1}{2\tau^2} \end{pmatrix} \tag{9}$$

where $\boldsymbol{\Phi}(\boldsymbol{\eta}_q)$ is a $(q \times q)$ matrix with entries $\Phi_{i,j}(\boldsymbol{\eta}_q) = \mathbb{E}\left[x_t x_{t+|i-j|}\right]$, which do not depend on the innovation variance τ. This is expected, given that the signal-to-noise ratio of a moving average model is also independent of τ. The asymptotic information matrix for a moving average model is therefore equivalent to the asymptotic information matrix for an autoregressive model with coefficients $\boldsymbol{\eta}_q$. This implies that $|\mathbf{J}(\boldsymbol{\eta}_q)| \geq n^q$ for all $\boldsymbol{\eta}_q \in \Lambda_q$, and that $|\mathbf{J}(\boldsymbol{\eta}_q)| \to \infty$ as the coefficients approach the boundary of the invertibility region. The entries of the autocovariance matrix $\boldsymbol{\Phi}(\boldsymbol{\eta}_q)$ can be computed using the formulae presented in [9]; however, in Section 2.2, a simplified expression for the message length is presented in which there is no need to explicitly compute the autoregressive autocovariance matrix. This results in significant increases in both numerical stability and computational efficiency.

2.2 Minimising the Message Length

The minimum message length estimates are the values of the parameters $\boldsymbol{\theta}_q$ that minimise the MML87 message length. Due to the difficulty in maximising the likelihood, these estimates must be found by a numerical search. The invertibility region, Λ_q, which defines the set of permissible moving average coefficients, forms a complex polyhedron for $q \geq 3$, making a constrained numerical search difficult. An alternative, and convenient, reparameterisation is in terms of reflection coefficients or partial autocorrelations. There exists a one-to-one transformation between partial autocorrelations, $\boldsymbol{\rho}$, in the invertibility region, and moving average coefficients $\boldsymbol{\eta}_q(\boldsymbol{\rho})$ [15]. In partial autocorrelation space the invertibility region, P_q, reduces to the interior of a hyper-cube

$$P_q = \{\boldsymbol{\rho} : |\rho_j| < 1, j = 1, \ldots, q\},$$

considerably simplifying the constrained minimisation problem. A further benefit to performing the numerical minimisation in partial autocorrelation space is that the determinant of the *coefficient-space* asymptotic Fisher information matrix, (9), is given by the simple expression

$$|\mathbf{J}(\boldsymbol{\theta}_q)| = \left(\frac{n^{q+1}}{2\tau^2}\right)\left(\prod_{j=1}^{q} \frac{1}{(1-\rho_j^2(\boldsymbol{\eta}_q))^j}\right), \tag{10}$$

where $\boldsymbol{\rho}(\boldsymbol{\eta}_q) = (\rho_1(\boldsymbol{\eta}_q), \ldots, \rho_q(\boldsymbol{\eta}_q))'$ are the q partial autocorrelations corresponding to the coefficients $\boldsymbol{\eta}_q$. In contrast to direct evaluation of $|\mathbf{J}(\boldsymbol{\theta}_q)|$ in coefficient space, this expression involves only $O(q)$ operations and does not require the direct computation of the autocovariances of the auxilliary autoregressive process (8). Using (6), (7) and (10) in (3) yields the following expression for the MML87 message length, $I_{87}(\mathbf{y}, \boldsymbol{\eta}_q, \tau, q)$,

$$\begin{aligned} -\log p_q(\mathbf{y}|\boldsymbol{\eta}_q, \tau) + \frac{q}{2}\log n - \frac{1}{2}\sum_{j=1}^{q} j \log(1-\rho_j^2(\boldsymbol{\eta}_q)) + \log \operatorname{vol}(\Lambda_q) + c(q+1) \\ + \frac{1}{2}\log\left(\frac{n}{2}\right) + \log\left[(Q+1)\log\left(\frac{\tau_1}{\tau_0}\right)\right], \end{aligned} \tag{11}$$

where $c(\cdot)$ is given by (4). The last two terms of (11) are constant with respect to q and $\boldsymbol{\theta}_q$, and therefore have no effect on parameter estimation or order selection, and may be ignored if we are only considering moving average models to be possible explanations of the data. The above expression easily handles the case that $q = 0$ by simply dropping the second through fourth terms. It is important to note that (11) gives an expression for the message length in terms of the coefficients $\boldsymbol{\eta}_q$; the corresponding partial autocorrelations, $\boldsymbol{\rho}(\boldsymbol{\eta}_q)$, are only used because they make evaluating and minimising this expression significantly easier. The MML87 estimate of the innovation variance, $\hat{\tau}^{87}(\boldsymbol{\eta}_q)$, conditional on a coefficient vector $\boldsymbol{\eta}_q$, is the same as the maximum likelihood estimate,

$$\hat{\tau}^{87}(\boldsymbol{\eta}_q) = \frac{\mathbf{y}'\boldsymbol{\Gamma}(\boldsymbol{\eta}_q)\mathbf{y}}{n},$$

which itself may be calculated efficiently through the use of the same Kalman filter recurrence relations used to calculate the negative log-likelihood. The MML87 parameter estimates, $\hat{\boldsymbol{\eta}}_q^{87}$, are found by searching for the partial autocorrelations that solve

$$\hat{\boldsymbol{\rho}}^{87} = \underset{\boldsymbol{\rho} \in P_q}{\arg\min} \left\{ I_{87}(\mathbf{y}, \boldsymbol{\eta}_q(\boldsymbol{\rho}), \hat{\tau}^{87}(\boldsymbol{\eta}_q(\boldsymbol{\rho})), q) \right\},$$

and transforming them to coefficient space, i.e., $\hat{\boldsymbol{\eta}}_q^{87} \equiv \boldsymbol{\eta}_q(\hat{\boldsymbol{\rho}}^{87})$. The MML estimate of the order, $\hat{q}^{87}$, may then be found by solving

$$\hat{q}^{87} = \underset{q \in \{0,\ldots,Q\}}{\arg\min} \left\{ I_{87}(\mathbf{y}, \hat{\boldsymbol{\eta}}_q^{87}, \hat{\tau}^{87}(\hat{\boldsymbol{\eta}}_q^{87}), q) \right\}.$$

An interesting result is that evaluation of the MML87 message length (11) involves only $O(q)$ additional operations over evaluation of the negative log-likelihood, and thus the minimum message length estimates are theoretically as quick to find numerically as the maximum likelihood estimates. In fact, experiments suggest that the extra "regularisation" introduced by the presence of the asymptotic Fisher information term acts to significantly improve convergence of the search procedure for MML estimates in comparison to maximum likelihood estimation, which is well known to be problematic for moving average models.

2.3 Properties of the MML87 Estimator

The MML87 estimate of order, $\hat{q}^{87}$, is a strongly consistent estimate of q^*. To see this, rewrite the message length (11) as

$$-\log p_q(\mathbf{y}|\boldsymbol{\eta}_q, \tau) + \frac{q}{2}\log n + O(1),$$

where $O(1)$ denotes terms that are constant with respect to n. Thus, the MML87 message length (11) asymptotically coincides with the Bayesian information criterion (BIC), and from the arguments in [16], $\hat{q}^{87}$ is a strongly consistent estimate of q^*. Further, the MML87 estimates of the coefficients and innovation variance asymptotically coincide with the maximum likelihood estimates. This implies that when $q \geq q^*$ the MML87 parameter estimates are also strongly consistent [17]. The MML87 estimates of the moving average coefficients also possess an interesting finite sample property.

Property 1. *For all datasets,* $\mathbf{y}$*, of all finite sample sizes* n*, the partial autocorrelations corresponding to the MML87 estimates of the coefficients,* $\hat{\boldsymbol{\eta}}^{87}$*, satisfy*

$$||\boldsymbol{\rho}(\hat{\boldsymbol{\eta}}^{87})||_\infty < 1,$$

where $||\cdot||_\infty$ *denotes the* ℓ_∞ *norm.*

Proof. The MML estimates minimise the sum of the negative log-likelihood and the half log-determinant of the Fisher information matrix, as given in (11). The negative log-likelihood is bounded from below for finite n for all $\boldsymbol{\eta}_q \in \mathbb{R}^q$; in contrast, the half log-determinant of the Fisher information matrix is unbounded from above as $||\boldsymbol{\rho}(\boldsymbol{\eta}_q)||_\infty \to 1$. Thus, the message length will be finite if and only if $||\boldsymbol{\rho}(\hat{\boldsymbol{\eta}}^{87})||_\infty < 1$, implying that the parameter estimates that minimise (11) must satisfy $||\boldsymbol{\rho}(\hat{\boldsymbol{\eta}}^{87})||_\infty < 1$. □

This result implies that $\hat{\boldsymbol{\eta}}^{87} \in \mathit{\Lambda}_q$, and therefore the MML87 estimates of the moving average coefficients do not suffer from the so-called "pile-up" phenomenon [18], in which coefficients are estimated to lie exactly on the boundary of the invertibility region; this problem is well known to affect the maximum likelihood estimates. The removal of the troublesome pile-up effect is attributable to the "regularisation" introduced by the Fisher information terms, which also corroborates the empirical observations that the message length surface is better behaved than the likelihood surface when performing numerical optimisation.

3 Evaluation

Two measures of "closeness" to the true, generating moving average process were used to assess the competing estimators: (i) normalized expected one-step-ahead squared prediction error; and (ii) the directed Kullback–Leibler divergence. The expected one-step-ahead squared prediction error is defined as the expected squared difference between the true conditional mean and the predicted conditional mean for the next sample if the q previous innovations were available, and assesses the closeness of the estimated moving average model in ideal conditions. Given a true model, $\boldsymbol{\eta}^*$, and estimated model, $\hat{\boldsymbol{\eta}}$, this is equal to

$$\mathrm{SPE}_1(\boldsymbol{\eta}^*, \hat{\boldsymbol{\eta}}) = \left(\frac{1}{\gamma_0(\boldsymbol{\eta}^*)}\right)(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}})'(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}}), \tag{12}$$

where $\gamma_0(\boldsymbol{\eta}^*)$ is the zero-order autocovariance of the generating process (found using (5)), and the two parameter vectors are made to be the same dimension by appending a suitable number of zero elements to the shorter vector. The scaling by the inverse of the zero-order autocovariance of the generating process renders the resulting quantity unitless, and is done to ensure that the value of the error metric is comparable between different generating processes. This is essential if simulations involve sampling a large number of generating models from the invertibility region, with correspondingly different signal-to-noise ratios. The second error metric used was the Kullback–Leibler (K–L) divergence [19]. This is an important, parameterisation-invariant measure of the "distance" between distributions, with strong information theoretic interpretations. The *per sample* K–L divergence between a true, generating moving average process, $\boldsymbol{\theta}^*$, and an approximating moving average process, $\hat{\boldsymbol{\theta}}$, for n data points is

$$\frac{1}{2}\log\left(\frac{\hat{\tau}}{\tau^*}\right) + \left(\frac{1}{2n}\right)\log\left(\frac{|\boldsymbol{\Gamma}(\hat{\boldsymbol{\eta}}|}{|\boldsymbol{\Gamma}(\boldsymbol{\eta}^*)|}\right) + \left(\frac{1}{2n}\right)\mathrm{Tr}\left(\boldsymbol{\Gamma}(\boldsymbol{\eta}^*)\boldsymbol{\Gamma}^{-1}(\hat{\boldsymbol{\eta}})\right) - \frac{1}{2}, \tag{13}$$

where $\boldsymbol{\Gamma}(\cdot)$ is an $(n \times n)$ autocovariance matrix as in (5). The choice of the size of the autocovariance matrix is essentially arbitrary and the use of the sample size,

n, reflects the fact that the sequence $\mathbf{y}$ can be regarded as a randomly generated vector from the n-dimensional multivariate normal distribution characterised by $\tau^* \boldsymbol{\Gamma}(\boldsymbol{\eta}^*)$. Thus, the Kullback–Leibler divergence measures the closeness of the estimated n-dimensional multivariate distribution to the distribution that generated the sample.

3.1 Parameter Estimation

The parameter estimation performance of the MML87 estimator was compared against two standard procedures from the literature: (i) the maximum likelihood (ML) estimator, and (ii) the modified Durbin estimator (ARMASA) [20, 21] which exploits the duality between moving average and autoregressive models. Given our choice of a uniform prior distribution for the moving average coefficients, the maximum likelihood estimator also coincides with the maximum a posteriori (MAP) estimator. As we know that the mean of the generating moving average process is zero, we chose not to demean the data before estimation by ARMASA to allow for a fair comparison with MML87 and ML.

The simulation setup was as follows: (i) sample an invertible moving average model $\boldsymbol{\eta}_{q^*}$ uniformly from Λ_{q^*} (using the algorithm described in [22]); (ii) sample a time series of length n from the process defined by $\boldsymbol{\eta}_q^*$, with $\tau^* = 1$; (iii) estimate coefficients from the time series using MML87, maximum likelihood, the ARMASA procedure, and compute appropriate measures of closeness to the generating model. This was repeated for 10^3 iterations, for $q = \{1, 4, 7, 10\}$, with sample sizes $n = k(3q + 1)$, where $k = \{1, 2, 4\}$.

The results are presented in Table 1. Median expected one-step-ahead squared error, (12), and median per sample Kullback–Leibler divergences, (13), were used instead of arithmetic means as the tails of the empirical distributions of these error measures were significantly heavier than would be expected for a normal distribution. The results clearly show the strong performance of the MML87 estimator, which is superior in terms of both squared prediction errors, and Kullback–Leibler divergence, for every combination of sample size and true order. For the smallest sample sizes the maximum likelihood/MAP estimator uniformly performed the worst, often by a large margin; this can be attributed to the “pile-up” effect as well as the tendency of the maximum likelihood estimator to overestimate the magnitude of the zeros of the underlying process when the sample size is small. In contrast, observations suggested that the MML87 estimates tended to underestimate the magnitudes of the zeros in comparison to the maximum likelihood estimates, and the distributions of the estimates appeared unimodal, showing no sign of any “pile-up” type effect. For larger sample sizes, the modified Durbin’s method generally performed worse than the maximum likelihood estimator in terms of squared errors. Of course, the true models used in the simulations have been sampled from the prior distribution used by the MML87 estimator, which makes direct comparisons with maximum likelihood and the modified Durbin’s method somewhat problematic. However, even taking this into account, the results demonstrate that the MML87 estimator is clearly

Order	n	Squared Prediction Error			Kullback–Leibler Divergence		
		MML87	ML	ARMASA	MML87	ML	ARMASA
1	4	**0·071**	0·143	0·085	**0·139**	0·182	0·152
	8	**0·032**	0·055	0·051	**0·089**	0·105	0·090
	16	**0·015**	0·022	0·025	**0·040**	0·048	0·048
4	13	**0·158**	0·297	0·238	**0·182**	0·313	0·236
	26	**0·070**	0·102	0·111	**0·096**	0·147	0·120
	52	**0·031**	0·038	0·053	**0·048**	0·063	0·066
7	22	**0·164**	0·320	0·266	**0·210**	0·382	0·261
	44	**0·077**	0·116	0·126	**0·108**	0·173	0·132
	88	**0·033**	0·041	0·058	**0·052**	0·070	0·068
10	31	**0·172**	0·315	0·294	**0·209**	0·390	0·278
	62	**0·079**	0·117	0·123	**0·111**	0·185	0·133
	124	**0·035**	0·043	0·058	**0·051**	0·075	0·069

Table 1. Parameter estimation experiment results.

superior to the usual Bayesian MAP estimator which utilises the same prior information.

One point of particular note is that despite the fact that the message length formula (11) is based on the *asymptotic* Fisher information matrix, the MML87 estimates perform very well in the small sample regime. This suggests that further refinements of the message length formula to make use of the finite sample Kullback–Leibler divergence, such as the new message length formulae discussed in [5] (pp. 30–34) could lead to further performance improvements for small samples. This issue, along with a more complete characterisation of the behaviour of the MML87 moving average estimates, are interesting topics for future research.

3.2 Order Selection

The ability of the MML87 criterion to estimate a moving average model of unknown order from finite samples was compared against six standard procedures from the literature: the Akaike information criterion (AIC) [1], the corrected AIC (AIC$_c$) [23], the symmetric Kullback–Leibler divergence criterion (KIC) [24], the corrected KIC (KIC$_c$) [25] and the ARMAsel procedure [26]. The BIC, AIC, KIC and their corrected variants use maximum likelihood estimates, while the ARMAsel procedure uses modified Durbin estimates (without zero meaning the data, as previously discussed).

The simulation setup was as follows: (i) sample an invertible moving average model $\boldsymbol{\eta}_{q^*}$ uniformly from Λ_{q^*}; (ii) sample a time series of length n from the process defined by $\boldsymbol{\eta}^*_q$, with $\tau^* = 1$; (iii) ask all criteria to estimate q^* along with estimates of the moving average coefficients, and compute appropriate measures of closeness to the generating model. This was repeated one thousand times for each true model order $q^* = \{0, \ldots, 10\}$, for a total of $11,000$ iterations per sample size $n = \{10, 20, 50, 100\}$. At each iteration, all candidate models in $q = \{0, \ldots, r\}$ were considered by the model selection criteria, with $r = 4$ for

n	Measure	Model Selection Criteria						
		MML87	BIC	AIC	AIC$_c$	KIC	KIC$_c$	ARMASA
	SPE$_1$	**0·387**	0·412	0·414	0·426	0·414	0·444	0·405
10	KL	**0·220**	0·247	0·258	0·246	0·242	0·255	0·239
	$\#\{\hat{q} = q^*\}$	1534	1607	**1670**	1390	1499	1250	1481
	SPE$_1$	**0·245**	0·284	0·292	0·287	0·284	0·309	0·268
20	KL	**0·175**	0·206	0·236	0·206	0·206	0·207	0·186
	$\#\{\hat{q} = q^*\}$	2341	2292	**2487**	2186	2292	2032	2243
	SPE$_1$	**0·065**	0·093	0·109	0·091	0·086	0·090	0·086
50	KL	**0·084**	0·114	0·138	0·117	0·111	0·113	0·094
	$\#\{\hat{q} = q^*\}$	**4343**	3912	4235	4281	**4343**	3978	4192
	SPE$_1$	**0·022**	0·028	0·037	0·033	0·028	0·027	0·036
100	KL	**0·036**	0·047	0·056	0·052	0·046	0·045	0·047
	$\#\{\hat{q} = q^*\}$	**6227**	5799	5584	5817	6200	6178	5997

Table 2. Order estimation experiment results.

$n = 10$, $r = 7$ for $n = 20$, and $r = 10$ for $n > 20$. These simulations were designed to mimic real data situations in which the true, underlying process may be considerably more complex than any model that the data will allow us to realistically consider.

The median expected one-step-ahead squared prediction error, (12), median per sample Kullback–Leibler divergence, (13), and the number of times a criteria correctly estimated the true model order are presented in Table 2. In terms of squared prediction errors, and Kullback–Leibler divergence, the MML87 criterion is uniformly the best for all sample sizes. For $n = 10$ and $n = 20$, the ARMAsel procedure performed similar to MML87, although for $n > 20$ it performed noticeably poorer. In terms of correct order selections, for all but $n = 50$ the AIC and/or KIC perform amongst the best of all the methods. Not too much should be made of this fact, however, as it is well known that as n grows these criteria are inconsistent and will tend to overfit with non-vanishing probability. Interestingly, for smaller sample sizes, despite performing the best in terms of predictive measures (squared error and Kullback–Leibler divergence) the MML87 criterion does not, in general, perform the best at selecting the true generating order. We believe this is because MML, and related compression based methods, make no assumptions about the existence of a “true” model; rather, they are designed to select a good, *plausible* explanation about the data generating source from the available candidates.

3.3 The Southern Oscillation Index Time Series

Finally, we conclude this section with a brief experiment on a real time series. The Southern Oscillation Index (SOI) is a time series of monthly measurements of fluctuations in air pressure difference between Tahiti and Darwin. The SOI is commonly used to study and predict *El Niño* phenomena. The time series analysed contained $n = 1,619$ monthly measurements taken from January, 1876 through to December, 2010, and was obtained from the Australian Government

Measure	Model Selection Criteria						
	MML87	BIC	AIC	AIC$_c$	KIC	KIC$_c$	ARMASA
SPE$_1$	**59·129**	59·166	59·740	59·740	59·740	59·740	59·792
NLL	**2490·1**	2490·5	2492·3	2492·3	2492·3	2492·3	2491·4
Order	13	13	14	14	14	14	14

Table 3. Southern Oscillation Index Experiment.

Bureau of Metereology website. The first $1,000$ samples in the series were used as a training sample, and the remaining 619 samples were used for validation of the estimated models. All criteria were asked to estimate a suitable moving average model from the training data, with a maximum candidate order of $q = 20$. The evaluation measures were mean squared prediction error and negative log-likelihood obtained on the validation sample, conditional on the training sample. These were computed by running the Kalman filter on the complete time series using the models estimated from the training sample; this automatically produces predictions of the mean and variance for all data points in the time series, and these may be used to compute the squared error and negative log-likelihood of the validation sample (i.e., the last 619 samples).

The mean squared error and negative log-likelihood scores are presented in Table 3, along with the order of the moving average model selected by each of the criteria. All criteria perform similarly, with MML87 obtaining a slight improvement in squared prediction error. The two main points of interest are: (i) even for this large sample ($n = 1,000$), the MML87 criteria has selected a slightly lower order model ($\hat{q}^{87} = 13$) than all other criteria except BIC; and (ii) the MML87 estimates of coefficients for the $q = 13$ moving average model still differ slightly from the maximum likelihood estimates used by BIC, despite the high ratio of data-to-parameters.

Appendix A

For completeness we present the equations (taken from [12]) to calculate the volume of the invertibility region, Λ_q, for a q-th order moving average process. Define $M_1 = 2$ and $M_k = ((k-1)/k)M_{k-2}$. Let $V_q \equiv \mathrm{vol}(\Lambda_q)$, with $V_1 = 2$; for $q > 1$, $V_q = \prod_{k=0}^{q/2-1} M_{2k+1}^2$ for q even and $V_q = V_{q-1}M_q$ for q odd.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6) (December 1974) 716–723
2. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2) (1978) 461–464
3. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
4. Sak, M., Dowe, D., Ray, S.: Minimum message length moving average time series data mining. In: Proceedings of the ICSC Congress on Computational Intelligence Methods and Applications (ACFM2005), Istanbul, Turkey (2005)

5. Schmidt, D.F.: Minimum Message Length Inference of Autoregressive Moving Average Models. PhD thesis, Clayton School of Information Technology, Monash University (2008)
6. Solomonoff, R.J.: A formal theory of inductive inference. Information and Control **7**(2) (1964) 1–22, 224–254
7. Farr, G.E., Wallace, C.S.: The complexity of strict minimum message length inference. Computer Journal **45**(3) (2002) 285–292
8. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) **49**(3) (1987) 240–252
9. Porat, B., Friedlander, B.: Computation of the exact information matrix of Gaussian time series with stationary random components. IEEE Transactions on Acoustics, Speech and Signal Processing **34**(1) (1986) 118–130
10. Gardner, G., Harvey, A.C., Phillips, G.D.A.: Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. Applied Statistics **29**(3) (1980) 311–322
11. Fitzgibbon, L.J., Dowe, D.L., Vahid, F.: Minimum message length autoregressive model order selection. In: Proceedings of the International Conference on Intelligent Sensing and Information Processing (ICISIP). (2004) 439–444
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. Makalic, E., Schmidt, D.F.: Fast computation of the Kullback-Leibler divergence and exact Fisher information for the first-order moving average model. IEEE Signal Processing Letters **17**(4) (2009) 391–393
14. Whittle, P.: The analysis of multiple stationary time series. Journal of the Royal Statistical Society, Series B (Methodological) **15**(1) (1953) 125–139
15. Barndorff-Nielsen, O., Schou, G.: On the parametrization of autoregressive models by partial autocorrelations. Journal of Multivariate Analysis **3** (1973) 408–419
16. Haughton, D.M.A.: On the choice of a model to fit data from an exponential family. The Annals of Statistics **16**(1) (March 1988) 342–355
17. Rissanen, J., Caines, P.E.: The strong consistency of maximum likelihood estimators for ARMA processes. The Annals of Statistics **7**(2) (1979) 297–315
18. Davidson, J.E.H.: Problems with the estimation of moving average processes. Journal of Econometrics **16**(3) (1981) 295–310
19. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1) (March 1951) 79–86
20. Durbin, J.: Efficient estimation of parameters in moving-average models. Biometrika **46**(3/4) (1959) 306–316
21. Broersen, P.M.T.: Autoregressive model orders for Durbin's MA and ARMA estimators. IEEE Transactions on Signal Processing **48**(8) (2000) 2454–2457
22. Jones, M.C.: Randomly choosing parameters from the stationarity and invertibility regions of autoregressive-moving average models. Applied Statistics **36**(2) (1987) 134–138
23. Hurvich, C.M., Tsai, C.L.: Regression and time series model selection in small samples. Biometrika **76**(2) (June 1989) 297–307
24. Cavanaugh, J.E.: A large-sample model selection criterion based on Kullback's symmetric divergence. Statistics & Probability Letters **42**(4) (1999) 333–343
25. Seghouane, A.K., Bekara, M.: A small sample model selection criterion based on Kullback's symmetric divergence. IEEE Transactions on Signal Processing **52**(12) (December 2004) 3314–3323
26. Broersen, P.M.T.: Automatic spectral analysis with time series models. IEEE Transactions on Instrumentation and Measurement **51**(2) (2002) 211–216