MONASH

WP 6/96

ISSN 1032-3813
ISBN 0 7326 0787 6

MONASH UNIVERSITY

ANCORA IMPARO

AUSTRALIA

ESTIMATION OF REGRESSION DISTURBANCES BASED ON MINIMUM MESSAGE LENGTH

Mizan R. Laskar and Maxwell L. King

Working Paper 6/96
June 1996

DEPARTMENT OF ECONOMETRICS

ESTIMATION OF REGRESSION DISTURBANCES BASED ON MINIMUM MESSAGE LENGTH

Mizan R. Laskar and Maxwell L. King*
Department of Econometrics
Monash University
Clayton, Victoria 3168
Australia

Abstract

This paper derives six different forms of message length functions for the general linear regression model. In so doing, two different prior densities and the idea of parameter orthogonality are employed. Parameter estimates are then obtained by finding those parameter values which minimize the message length. The asymptotic properties of the minimum message length (MML) estimators are studied and it is shown that these estimators are asymptotically normal. A Monte Carlo experiment was conducted to investigate the small sample properties of the MML estimators in the context of first-order moving average regression disturbances. The results show that the combination of parameter orthogonality and message length based inference can produce good small sample properties.

* This research has been supported in part by ARC grant.

1 Introduction

Estimation of the parameters involved in the variance-covariance matrix of linear regression disturbances has long been a problem in econometrics. This is because of the non-experimental nature of economic data. Econometric models usually involve a large number of influences, all of which are not of direct interest. As a result, nuisance parameters often need to be dealt with, because their presence can cause biases in estimates and tests of the parameter of interest.

Cox and Reid (1987) observed that estimates and tests of the parameter of interest based on the classical likelihood can give biased and inefficient results in small samples. As an alternative to the classical likelihood, there are a range of other likelihoods which deal with nuisance parameters in a more satisfactory manner. Recently, Laskar and King (1995) investigated a number of modified likelihoods and concluded that these can help remove the effects of nuisance parameters. They investigated the small sample properties of estimators in the context of first-order moving average (MA(1)) regression disturbances and reported a significant improvement in estimators based on modified likelihoods compared with their counterparts which are based on the classical likelihood.

Cox and Reid (1987) initiated the idea of the conditional profile likelihood, which essentially requires the orthogonality of the parameter of interest and nuisance parameters. A slightly different idea, known as minimum message length (MML) estimation, was introduced by Wallace and Boulton (1968), Boulton and Wallace (1970, 1973) and Boulton (1975) while working with the problem of classification. However, they mainly discussed a computer based method called SNOB. Also, Wallace and Freeman (1987) extended the idea of MML estimation from a Bayesian viewpoint as an alternative method of estimation and test construction for the parameters of interest. MML estimation is a Bayesian method which chooses estimates to minimize the length of a certain encoded form of the data, while maximum likelihood estimation is one which chooses estimates to maximize the likelihood function. Extending this research, Wallace and Freeman (1992) applied the MML approach to the problem of estimating the parameters of a multivariate Gaussian model and found that the MML estimates on average are more accurate than those of the maximum likelihood estimator. Following from this, Wallace and Dowe (1993) applied the MML approach to estimating the von Mises concentration parameter and observed its improved accuracy over the maximum likelihood estimator for small sample sizes. However the MML principle needs a prior distribution of the parameters and square root of the determinant of the Fisher information matrix for the parameters. In this context, Wallace and Dowe (1993) showed that the inclusion of these two factors helps reduce the measure of uncertainty.

In this paper, two MML estimators, denoted by MML_1 and MML_2 are derived for the general linear regression model with non-spherical disturbances, using two different prior distributions for the parameters. They are based on two message length formulae, ML_1 and ML_2, which contain the nuisance parameters, regressors and parameters of interest. The nuisance parameters may cause problems for estimation and tests of the parameters of interest. To overcome these problems, four further estimators are developed. Two of the estimators are obtained by minimizing the message length functions which are constructed using Cox and Reid's (1987) idea applied to ML_1 and ML_2. These two estimators are called $CMML_1$ and $CMML_2$. The remaining two estimators are obtained by minimizing the message length functions which are the combination of parameter orthogonality and message length, known as $AMML_1$ and $AMML_2$. The MML estimation technique and the asymptotic properties of the resultant estimates are also studied.

This paper is divided into a further four sections. All the different message length formulae are derived in Section 2. The properties of MML estimators are discussed in Section 3. A Monte Carlo experiment is conducted in Section 4 to investigate the small sample properties of estimators for MA(1) regression disturbances. Some concluding remarks are made in the final section.

2 Theory

Consider the general linear regression model

$$y = X\beta + u\,,\ u \sim N(0, \sigma^2 \Omega(\theta))\,, \tag{1}$$

where y is $n{\times}1$, X is $n{\times}k$, nonstochastic and of rank $k < n$, and $\Omega(\theta)$ is a symmetric matrix that is positive definite for the unknown p-vector θ belonging to a subspace of R^p. This model generalizes a wide range of disturbance processes of the linear regression model of particular interest to econometricians and statisticians. These include all parametric forms of autocorrelated disturbances, all parametric forms of heteroscedasticity (in which case $\Omega(\theta)$ is a diagonal matrix), and error components model including those that result from random regression coefficients. The log likelihood function for model (1) is

$$l(y;\theta,\beta,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}(y - X\beta)'\Omega(\theta)^{-1}(y - X\beta)/\sigma^2\,. \tag{2}$$

2.1 Derivation of the Message Length

For model (1), the message length is given by

$$-\log\left[\frac{\pi(\theta,\beta,\sigma^2)L(y;\theta,\beta,\sigma^2)}{\sqrt{F(\theta,\beta,\sigma^2)}}\right] + \frac{D}{2}(1 + \log K_D) \tag{3}$$

where $\pi(\theta,\beta,\sigma^2)$ is the prior density for $\gamma = (\theta',\beta',\sigma^2)$, $L(y;\theta,\beta,\sigma^2)$ is the likelihood of (1), $F(\theta,\beta,\sigma^2)$ is the determinant of the Fisher information matrix, D is the number of parameters and K_D is the D-dimensional lattice constant which is independent of parameters, as given by Conway et al. (1988). For model (1), the Fisher information matrix is given by

$$E\left(-\frac{\partial^2 l}{\partial\gamma\partial\gamma'}\right)=\begin{bmatrix} A(\theta) & B(\theta) & 0 \\ B'(\theta) & \frac{n}{2\sigma^4} & 0 \\ 0 & 0 & \frac{X'\Omega(\theta)^{-1}X}{\sigma^2} \end{bmatrix}$$

where the $(i,j)^{\text{th}}$ element of the $p\times p$ matrix $A(\theta)$ is $\frac{1}{2}tr\left[-\frac{\partial\Omega(\theta)}{\partial\theta_i}\frac{\partial\Omega(\theta)^{-1}}{\partial\theta_j}\right]$, the i^{th} element of the $p\times 1$ vector $B(\theta)$ is $\frac{1}{2\sigma^2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta_i}\right]$ and tr represents the trace of a matrix. For simplicity at this stage, we will assume θ is a scalar. The construction of message length for model (1) needs the determinant of the information matrix, log of which is

$$\log F(\theta,\beta,\sigma^2) = -(k+2)\log\sigma^2 + \log\left|X'\Omega(\theta)^{-1}X\right| + \log(n\times tr\left[-\frac{\partial\Omega(\theta)}{\partial\theta}\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2) - \log 4.$$

Assuming non-informative prior $\pi(\theta,\beta,\sigma^2) = \frac{1}{\sigma}$, the message length given by (3) is

$$\text{ML}_1 = \frac{m-1}{2}\log\sigma^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma^2}u'\Omega(\theta)^{-1}u + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| + \frac{1}{2}\log(n\times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2) + \frac{D}{2}(1+\log K_D) - \log 2 \quad (4)$$

where $u = y - X\beta$ and $m = n - k$. Using the non-informative prior $\pi(\theta,\beta,\sigma^2) = \frac{1}{\sigma^2}$, the message length given by (3) is

$$\text{ML}_2 = \frac{m}{2}\log\sigma^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma^2}u'\Omega(\theta)^{-1}u + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| + \frac{1}{2}\log(n\times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2) + \frac{D}{2}(1+\log K_D) - \log 2. \quad (5)$$

2.2 Derivation of Cox and Reid's Conditional Profile Message Length

The two message length formulae given by (4) and (5) contain the nuisance parameters β and σ^2. Their presence may cause problems for estimators and tests of θ based on ML_1 and ML_2. Therefore, these potential problems need to be removed. One suggestion is to eliminate the effect of β and σ^2 using the idea of Cox and Reid (1987). Extending their research, Laskar and King (1995) constructed the conditional profile likelihood for model (1). To do this for (4) and (5), orthogonality of the parameters (β, θ) and (σ^2, θ) requires investigation. The parameters are orthogonal if

$$E\left(\frac{\partial^2 ML_1}{\partial\beta\partial\theta}\right) = i_{\beta,\theta} = 0 \text{ and } E\left(\frac{\partial^2 ML_1}{\partial\theta\partial\sigma^2}\right) = i_{\theta,\sigma^2} = 0.$$

In our case, $i_{\beta,\theta} = 0$ but

$$i_{\theta,\sigma^2} = \frac{1}{2\sigma^2} tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right] \text{ and } E\left(-\frac{\partial^2 ML_1}{\partial\sigma^4}\right) = i_{\sigma^2,\sigma^2} = \frac{n+k+1}{2\sigma^2}.$$

We therefore need to make a transformation $(\sigma^2, \theta) \rightarrow (\delta, \theta)$ so that δ and θ are orthogonal. This transformation is given by Laskar and King (1995), and is obtained by solving

$$-i_{\theta,\sigma^2} = i_{\sigma^2,\sigma^2}\frac{\partial\sigma^2}{\partial\theta},$$

$$\text{or } -\frac{1}{2\sigma^2} tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right] = \frac{n+k+1}{2\sigma^2}\frac{\partial\sigma^2}{\partial\theta},$$

or equivalently

$$\frac{\partial}{\partial\theta}\log\sigma^2 = \frac{\partial}{\partial\theta}\log|\Omega(\theta)|^{-\frac{1}{(n+k+1)}},$$

which has as one of its solutions

$$\sigma^2 = \delta|\Omega(\theta)|^{-\frac{1}{(n+k+1)}}$$

where δ is a constant. Using this transformation, (4) can be written without constant terms as

$$ML_{1t} = \frac{m-1}{2}\log\delta + \frac{k+1}{n+k+1}\log|\Omega(\theta)| + \frac{1}{2}\log|X_\theta' X_\theta| + \frac{1}{2\delta}u_\theta' u_\theta$$
$$+\frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2)$$

where $u_\theta = y_\theta - X_\theta\beta, X_\theta = G(\theta)^{-\frac{1}{2}}X,\ y_\theta = G(\theta)^{-\frac{1}{2}}y$ and $G(\theta) = \Omega(\theta) / |\Omega(\theta)|^{\frac{1}{(n+k+1)}}$. If we replace β and δ by their MML estimators, the concentrated (or profile) message length based on ML_1 becomes

$$ML_{1tc} = \frac{m-1}{2}\log\hat{\delta} + \frac{k+1}{n+k+1}\log|\Omega(\theta)| + \frac{1}{2}\log|X_\theta' X_\theta| + \frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right]$$
$$-\left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2) \tag{6}$$

where $\hat{\delta} = \hat{u}_\theta'\hat{u}_\theta / (n-k-1),\ \hat{u}_\theta = y_\theta - X_\theta\hat{\beta}_\theta$ and $\hat{\beta}_\theta = (X_\theta' X_\theta)^{-1}X_\theta' y_\theta$.

The construction of Cox and Reid's conditional profile message length for θ needs the concentrated message length (6) to be adjusted by the addition of the following term

$$\frac{1}{2}\log\left|\frac{\partial^2 ML_{1tc}}{\partial\gamma_1\partial\gamma_1'}\right|_{\gamma_1=\hat{\gamma}_1} = \frac{1}{2}\log|X_\theta' X_\theta| - \frac{k+2}{2}\log\hat{\delta} + \frac{1}{2}\log\frac{n-k-1}{2}$$

where $\gamma_1 = (\beta', \delta)'$. Thus the conditional profile message length for θ without the constant term is

$$\text{CPML}_1 = \frac{m-k-3}{2}\log\hat{\delta} + \frac{k+1}{n+k+1}\log|\Omega(\theta)| + \log|X_\theta' X_\theta|$$
$$+\frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2). \tag{7}$$

Similarly Cox and Reid's conditional profile message length for the model (5) becomes

$$\text{CPML}_2 = \frac{m-k-2}{2}\log\hat{\delta}_1 + \frac{k}{n+k}\log|\Omega(\theta)| + \log|X_\theta^{\prime *} X_\theta^*|$$
$$+\frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2) \qquad (8)$$

where $\hat{\delta}_1 = \hat{u}_\theta^{\prime *}\hat{u}_\theta^* / m$, $\hat{u}_\theta^* = y_\theta^* - X_\theta^* \hat{\beta}_\theta^*$, $\hat{\beta}_\theta^* = (X_\theta^{\prime *} X_\theta^*)^{-1} X_\theta^{\prime *} y_\theta^*$, $X_\theta^* = G_1(\theta)^{-\frac{1}{2}} X$,

$y_\theta^* = G_1(\theta)^{-\frac{1}{2}} y$ and $G_1(\theta) = \Omega(\theta) / |\Omega(\theta)|^{\frac{1}{(n+k)}}$.

2.3 An Alternative Derivation of Message Length Using Orthogonality

The message length functions ML_1 and ML_2 do not incorporate parameter orthogonality and this aspect may cause problems for the estimator and tests of the parameter θ. In this section, the message length function for model (1) is derived following application of an orthogonal transformation to the non-orthogonal parameters and assuming θ is a p×1 vector. It is clear that the parameters (θ_i, σ^2) are not orthogonal. Using the results of Laskar and King (1995), we can transform from (θ_i, σ^2) to (θ_i, δ_2) so that θ_i and δ_2 are orthogonal via $\sigma^2 = \delta_2 / |\Omega(\theta)|^{\frac{1}{n}}$. Using this transformation (2) becomes

$$l_t(y;\theta,\beta,\delta_2) = -\frac{n}{2}\log\delta_2 - \frac{1}{2\delta_2}(y - X\beta)' G_2(\theta)^{-1}(y - X\beta) \qquad (9)$$

where $G_2(\theta) = \Omega(\theta) / |\Omega(\theta)|^{\frac{1}{n}}$. The construction of the message length function needs the determinant of the Fisher information matrix of the parameters in (9), which is

$$\left| E\left(-\frac{\partial^2 l_t(y;\theta,\beta,\delta_2)}{\partial\gamma_2 \partial\gamma_2'}\right)\right| = \left|\frac{X_\theta^{\prime\dagger} X_\theta^\dagger}{\delta_2}\right| \frac{n}{2\delta_2^{\,2}} |D(\theta)|$$

where the $(i,j)^{th}$ element of the $p \times p$ matrix $D(\theta)$ is $\frac{1}{2} tr\left[\frac{\partial^2 G_2(\theta)^{-1}}{\partial\theta_i \partial\theta_j} G_2(\theta)\right]$, $\gamma_2 = (\theta', \beta', \delta_2)'$ and $X_\theta^\dagger = G_2(\theta)^{-\frac{1}{2}} X$. The derivation of message length differs for a different prior. Using (3), (9) and the non-informative prior $\pi(\theta,\beta,\delta)=\frac{1}{\sqrt{\delta_2}}$, the message length is

$$\mathrm{AML}_1 = \frac{m-1}{2}\log\delta_2 + \frac{1}{2\delta_2}u_\theta'^{\dagger}u_\theta^{\dagger} + \frac{1}{2}\log\left|X_\theta'^{\dagger}X_\theta^{\dagger}\right| + \frac{1}{2}\log|D(\theta)| \tag{10}$$

where $u_\theta^{\dagger} = y_\theta^{\dagger} - X_\theta^{\dagger}\beta$ and $y_\theta^{\dagger} = G_2(\theta)^{-\frac{1}{2}}y$. If we consider the prior $\pi(\theta,\beta,\delta_2)=\frac{1}{\delta_2}$, (10) changes to

$$\mathrm{AML}_2 = \frac{m}{2}\log\delta_2 + \frac{1}{2\delta_2}u_\theta'^{\dagger}u_\theta^{\dagger} + \frac{1}{2}\log\left|X_\theta'^{\dagger}X_\theta^{\dagger}\right| + \frac{1}{2}\log|D(\theta)|. \tag{11}$$

There are some similarities between these message length formulae and the likelihood functions investigated by Laskar and King (1995). Returning to the case where θ is a scalar, the MML estimates of β and σ^2 conditional on θ from (5) are

$\hat{\sigma}^2 = (y - X\hat{\beta})'\Omega(\theta)^{-1}(y - X\hat{\beta})/m = \hat{u}'\Omega(\theta)^{-1}\hat{u}/m$ and $\hat{\beta} = (X'\Omega(\theta)^{-1}X)^{-1}X'\Omega(\theta)^{-1}y$.

Putting these estimates in (5), it can be written without a constant term as

$$\frac{m}{2}\log\hat{\sigma}^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| + \frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2). \tag{12}$$

The form (12) is closer to the marginal likelihood for θ (Tunnicliffe Wilson (1989)). The only difference is due to an additional term, namely

$$\frac{1}{2}\log(n \times tr\left[-\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\frac{\partial\Omega(\theta)}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2).$$

For $p > 1$, AML_2 given by (11) can be written in terms of σ^2 as

$$\mathrm{AML}_2 = \frac{m}{2}\log\sigma^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2\sigma^2}u'\Omega(\theta)^{-1}u + \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right| + \frac{1}{2}\log|D(\theta)|. \tag{13}$$

The MML estimates of β and σ^2 conditional on θ from (13) are the same as those from (5). If we put these MML estimates of β and σ^2 in (13), it can be written as

$$\frac{m}{2}\log\hat{\sigma}^2+\frac{1}{2}\log|\Omega(\theta)|+\frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right|+\frac{1}{2}\log|D(\theta)|. \tag{14}$$

The form (14) is also closer to the marginal likelihood for θ. The difference is the additional term $D(\theta)$. For computational purposes, the $(i,j)^{th}$ element of $D(\theta)$ can be rewritten as

$$\frac{1}{2}\log tr[2G_2(\theta)^{-1}\frac{\partial G_2(\theta)}{\partial\theta_j}G_2(\theta)^{-1}\frac{\partial G_2(\theta)}{\partial\theta_i}-G_2(\theta)^{-1}\frac{\partial^2 G_2(\theta)}{\partial\theta_j\partial\theta_i}].$$

3 MML Estimation and Asymptotic Normality

The message length in (3) can be written as

$$ML(y;\gamma)=-\log\left[\frac{\pi(\gamma)}{F(\gamma)}L(y;\gamma)\right]. \tag{15}$$

The MML estimator of γ is that value of γ which minimizes (15) in such a way that $\pi(\gamma)\neq 0$ in the neighbourhood of the estimate of γ. On the other hand, the maximum likelihood estimate of γ is that value of γ which maximizes $L(y;\gamma)$. If $ML(y;\gamma)$ is a twice differentiable function of γ in its range, the MML estimate of γ is (assuming existence) given by roots of

$$ML'(y;\gamma)=\frac{\partial ML(y;\gamma)}{\partial\gamma}=0. \tag{16}$$

A sufficient condition that any of these values (say $\hat{\gamma}$) be a local minimum is that $ML''(y;\hat{\gamma})>0$. The $ML(y;\gamma)$ in (15) after division by n can be written as

$$\frac{1}{n}ML(y;\gamma)=-\frac{1}{n}\log\pi(\gamma)+\frac{1}{n}\log F(\gamma)-\frac{1}{n}\log L(y;\gamma). \tag{17}$$

To study the properties of MML estimator, some regularity conditions are needed. In this context, Godfrey (1988) and Ara (1995) discussed regularity conditions for classical likelihood based tests and marginal likelihood based tests respectively. But in this setting of MML estimation some additional conditions are needed. These are (i) $logF(\gamma)$ increases slower than n as a result, $\frac{1}{n} log\ F(\gamma) \to 0$ as $n \to \infty$ and (ii) $\frac{1}{n} log\ L(y;\gamma)$ is typically not zero as $n \to \infty$. Under the conditions (i) and (ii), $\frac{1}{n} ML(y;\gamma)$ is asymptotically equal to $-\frac{1}{n} log\ L(y;\gamma)$. Thus asymptotically we can write

$$\frac{\partial ML(y;\gamma)}{\partial\gamma} = -\frac{\partial \log L(y;\gamma)}{\partial\gamma},\ E\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right) = -E\left(\frac{\partial \log L(y;\gamma)}{\partial\gamma}\right) = 0$$

and

$$E\left(\frac{\partial^2 ML(y;\gamma)}{\partial\gamma\partial\gamma'}\right) = -E\left(\frac{\partial^2 \log L(y;\gamma)}{\partial\gamma\partial\gamma'}\right) = I(\gamma_0).$$

Using Taylor's theorem, we have

$$\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\hat{\gamma}} = \left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0} + \left(\frac{\partial^2 ML(y;\gamma)}{\partial\gamma\partial\gamma'}\right)_{\gamma^*}(\hat{\gamma}-\gamma_0) \tag{18}$$

where $\hat{\gamma}$ is the MML estimate of γ, γ_0 is the true value of γ and γ^* is a value between $\hat{\gamma}$ and γ_0. Since $\hat{\gamma}$ is a root of (16), (18) can be written in the form

$$-\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0} = \left(\frac{\partial^2 ML(y;\gamma)}{\partial\gamma\partial\gamma'}\right)_{\gamma^*}(\hat{\gamma}-\gamma_0). \tag{19}$$

Premultiplying (19) by $I(\gamma_0)^{-\frac{1}{2}}$ gives

$$-I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0} = I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial^2 ML(y;\gamma)}{\partial\gamma\partial\gamma'}\right)_{\gamma^*} I(\gamma_0)^{-\frac{1}{2}} I(\gamma_0)^{\frac{1}{2}}(\hat{\gamma}-\gamma_0). \tag{20}$$

The expression (20) can be rearranged as

$$\begin{aligned} I(\gamma_0)^{\frac{1}{2}}(\hat{\gamma}-\gamma_0) &= -\left[I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial^2 ML(y;\gamma)}{\partial\gamma\partial\gamma'}\right)_{\gamma^*} I(\gamma_0)^{-\frac{1}{2}}\right]^{-1} I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0} \\ &= I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0} + o_p(I) \end{aligned}$$

where $o_p(1)$ is the remainder term and converges in probability to zero (Ara (1995)). Hence $I(\gamma_0)^{\frac{1}{2}}(\hat{\gamma}-\gamma_0)$ is asymptotically equivalent to $I(\gamma_0)^{-\frac{1}{2}}\left(\frac{\partial ML(y;\gamma)}{\partial\gamma}\right)_{\gamma_0}$ therefore, these two expressions have the same asymptotic distribution, namely $N(0,I_s)$, under the null hypothesis of the true parameter value γ_0 (Crowder (1976), Godfrey (1988) and Ara (1995)), where $s = p+k+1$. Thus under the null hypothesis

$$I(\gamma_0)^{\frac{1}{2}}(\hat{\gamma}-\gamma_0) \rightarrow N(0,I_s).$$

So, $\hat{\gamma}$ is asymptotically normally distributed with mean vector γ_0 and variance-covariance matrix $I(\gamma_0)^{-1}$.

This result is not surprising if we view the MML estimator as a Bayesian estimator with prior $\pi(\gamma)/F(\gamma)$. In this context, Zellner (1983) mentioned that Bayesian estimators are consistent and normally distributed in large samples with asymptotic mean the same as that of the maximum likelihood estimator and asymptotic variance-covariance matrix equal to the inverse of the information matrix. Also, Heyde and Johnstone (1979) derived the asymptotic normality of the estimator based on the posterior distribution.

4 Monte Carlo Experiment

A Monte Carlo experiment was conducted to investigate the small sample properties of the six different MML estimators when the disturbances in (1) follow the MA(1) process

$$u_t = e_t + \theta e_{t-1} \tag{21}$$

with $e_t \sim IN(0,\sigma^2)$, $t = 0,1,\dots,n$. The classical likelihood function takes the same value when θ is replaced by $\frac{1}{\theta}$ and σ^2 by $\sigma^2\theta^2$ and therefore has a identification problem. The simplest solution to this problem is to estimate θ in the interval $-1 \leq \theta \leq 1$. For more details of this point see Laskar and King (1995). We simulated the estimator of θ in the context of (21) using each of the message length functions.

4.1 Experimental Design

The small sample properties of the six MML estimators and the classical maximum likelihood estimator of θ were investigated for each of the following $n \times k$ X matrices with n = 30 and $n = 60$.

X1: (k=5). A constant, quarterly Australian private capital movements, Government capital movements commencing 1968(1) and these two variables lagged one quarter as two additional regressors.

X2: (k=3). A constant, the first n observations of Durbin and Watson's (1951, p159) annual consumption of spirits example.

X3: (k=1). A constant.

X4: (k=4). A constant and three quarterly seasonal dummy variables.

X5: (k=2). A constant and a linear trend.

These matrices reflect different patterns. The *X1* matrix contains large volatile regressors, *X2* contains much less volatile annual regressors and *X5* contains the linear trend as a regressor. The values of θ used in this experiment were θ = -0.4, 0, 0.4 and 2000 replications were used throughout.

4.2 Empirical Results

Estimated bias, standard deviation, skewness and kurtosis of the seven different estimators of the MA(1) parameter for θ = 0, 0.4, -0.4 are presented in Table 1, 2, 3 respectively. The loss function, |bias| + $\frac{1}{\lambda}$(standard deviation) + $\frac{1}{\lambda^2}$|skewness| + $\frac{1}{\lambda^3}$|kurtosis - 3| where λ = 3, was used to summarize the estimated losses of the different estimators. These calculated losses are presented in Table 4.

The results reflect that for $\theta = 0$, estimates based on ML_1, ML_2, AML_1 and AML_2 have smaller biases compared to CML_1 and CML_2 but the estimates based on AML_1 and AML_2 have skewness closer to zero and kurtosis closer to 3 for all the X matrices. For $\theta = 0.4$ and -0.4, the estimates based on AML_1 and AML_2 have smaller bias and variance.

Estimated losses of the estimators based on the classical likelihood, CML_1 and CML_2 are larger compared to those of the other estimators. The losses of the estimators based on the classical likelihood are largest for most of the X matrices and those based on the AML_2 are smallest for most of the X matrices. The next smallest losses of estimators are for those based on AML_1. In general, estimates based on AML_1 and AML_2 have smaller biases and variances for $\theta = 0.4$, 0, -0.4 and are closer to the normal distribution. In contrast, estimates based on the classical likelihood, CML_1 and CML_2 have larger biases. It is clear that Cox and Reid's message length functions do not give relatively good estimates of θ. Estimated bias, standard deviation, skewness and kurtosis increase as θ moves closer to ± 1. As a result, losses of the estimators increase. This erratic behaviour occurs due to the identification problem as discussed above. It is clear that these findings are closer to the results obtained by Laskar and King (1995) where they pointed out that the distributions of the estimates of the MA(1) parameter are closer to the normal distribution for different modified likelihood functions, in contrast to the classical maximum likelihood estimator.

5 Conclusions

This paper has derived six different message length formulae, using two different prior densities for the parameters of the linear regression model with non-spherical error variance-covariance matrices. It has also investigated the asymptotic properties of the resultant MML estimators. It is observed that the asymptotic distribution of MML estimator is normal. Our simulation results show that Cox and Reid's modified CML_1 and CML_2 based estimates do not perform well. This may be because Cox and Reid's modification adds more information which, because of its nature, is already contained in the message length function. Estimates based on AML_2 are closer to normal when θ is closer to zero, which is significant improvement over the results reported by Laskar and King (1995).

Acknowledgements

We are grateful to an anonymous referee for constructive suggestions.

References

Ara, I. (1995), "Marginal likelihood based test of regression disturbances", unpublished Ph.D. thesis, Monash University.

Ara, I. and M.L. King (1995), "Marginal likelihood based tests of a subvector of the parameter vector of linear regression disturbances", *Proceedings of the Econometrics Conference at Monash University*, 69-106.

Boulton, D.M. (1975), "The information measure criterion for intrinsic classifications", Ph.D. thesis, Monash University.

Boulton, D.M. and C.S. Wallace (1970), "A programme for numerical classification", *Computational Journal* 13, 63-69.

Boulton, D.M. and C.S. Wallace (1973), "An information measure for hierarchic classification", *Computational Journal* 16, 245-261.

Conway, J.H. and N.J.A. Sloane (1988), *Sphere Packings, Lattice and Groups*, Spring-Verlag.

Cox, D.R. and N. Reid (1987), "Parameter orthogonality and approximate conditional inference (with discussion)", *Journal of the Royal Statistical Society* B 49, 1-39.

Crowder, M.J. (1976), "Maximum likelihood estimation for dependent observations", *Journal of Royal the Statistical Society* B 38, 45-53.

Durbin, J. and G.S. Watson (1951), "Testing for serial correlation in least squares regression II", *Biometrika* 38,159-178.

Godfrey, L.G. (1988), *"Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Related Approaches"*, Cambridge University Press.

Heyde, C.C. and I.M. Johnstone (1979), "On asymptotic posterior normality for stochastic process", *Journal of the Royal Statistical Society* B 41, 184-189.

Laskar, M.R. and M.L. King (1995), "Parameter orthogonality and likelihood functions", *Proceedings of the Econometrics Conference at Monash University*, 253-289.

Tunnicliffe Wilson, G. (1989), "On the use of marginal likelihood in time series model estimation", *Journal of the Royal Statistical Society* B 51, 15-27.

Wallace, C.S. and D.M. Boulton (1968), "An information measure classification", *Computational Journal* 11, 185-195.

Wallace, C.S. and D.L. Dowe (1993), "MML estimation of the von Mises concentration parameter", Technical Report No. 93/193, Monash University.

Wallace, C.S. and P.R. Freeman (1987), "Estimation and inference by compact coding", *Journal of the Royal Statistical Society* B 49, 240-265.

Wallace, C.S. and P.R. Freeman (1992), "Single-factor analysis by minimum message length estimation", *Journal of the Royal Statistical Society* B 54, 195-209.

Zellner, A. (1983), "Statistical theory and econometrics", in *Handbook of Econometrics*, eds, Z. Griliches and M.D. Intriligator, North Holland, 67-178.

Table 1. Estimated bias, standard deviation, skewness and kurtosis of the four estimators of the MA(1) parameter when θ = 0.0.

Design Matrix	Statistic	Estimator						
		MML						Max. Likeli. Est.
		ML_1	CML_1	ML_2	CML_2	AML_1	AM_2	
X1 n = 30	Bias	0.001	0.100	-0.002	0.095	0.006	0.003	-0.122
	S.D.	0.276	0.174	0.281	0.181	0.230	0.233	0.412
	Skewness	0.097	0.093	0.130	0.048	0.006	0.007	0.173
	Kurtosis	5.795	3.943	5.767	4.611	2.946	2.956	3.707
X1 n = 60	Bias	0.004	0.048	0.004	0.047	0.004	0.004	-0.040
	S.D.	0.151	0.134	0.151	0.134	0.149	0.149	0.183
	Skewness	0.001	0.053	0.000	0.050	0.001	0.002	0.563
	Kurtosis	4.339	4.278	4.348	4.270	3.605	3.616	7.159
X2 n = 30	Bias	0.006	0.170	0.000	0.161	0.008	0.004	-0.224
	S.D.	0.257	0.220	0.261	0.222	0.219	0.221	0.401
	Skewness	0.014	1.035	0.035	0.976	0.031	0.034	0.613
	Kurtosis	5.772	6.311	5.808	6.246	2.993	3.003	2.925
X2 n = 60	Bias	0.002	0.067	0.001	0.065	0.002	0.001	-0.064
	S.D.	0.148	0.134	0.148	0.134	0.146	0.146	0.173
	Skewness	0.034	0.015	0.038	0.014	0.003	0.004	0.569
	Kurtosis	4.282	3.307	4.311	3.310	3.394	3.404	6.136
X3 n = 30	Bias	0.007	0.059	0.005	0.056	0.009	0.007	-0.044
	S.D.	0.230	0.209	0.232	0.213	0.207	0.208	0.262
	Skewness	0.146	0.038	0.191	0.028	0.013	0.014	0.555
	Kurtosis	6.116	5.090	6.207	5.373	3.237	3.249	5.679
X3 n = 60	Bias	0.004	0.024	0.004	0.024	0.004	0.004	-0.015
	S.D.	0.140	0.137	0.140	0.137	0.140	0.140	0.174
	Skewness	0.000	0.001	0.000	0.001	0.000	0.000	0.007
	Kurtosis	3.415	3.350	3.417	3.353	3.365	3.367	3.531
X4 n = 30	Bias	0.003	0.005	0.004	0.004	0.008	0.009	0.005
	S.D.	0.250	0.222	0.252	0.227	0.211	0.212	0.276
	Skewness	0.331	0.248	0.305	0.355	0.010	0.010	0.157
	Kurtosis	6.799	6.202	6.784	6.471	3.201	3.216	6.501
X4 n = 60	Bias	0.004	0.004	0.004	0.004	0.004	0.004	0.004
	S.D.	0.144	0.142	0.144	0.142	0.143	0.143	0.148
	Skewness	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Kurtosis	3.552	3.430	3.555	3.435	3.479	3.482	3.715
X5 n = 30	Bias	0.007	0.116	0.004	0.111	0.009	0.006	-0.127
	S.D.	0.238	0.204	0.240	0.208	0.212	0.214	0.334
	Skewness	0.095	0.266	0.108	0.308	0.012	0.014	1.043
	Kurtosis	5.771	5.227	5.760	5.354	3.077	3.088	4.429
X5 n = 60	Bias	0.002	0.045	0.002	0.043	0.003	0.002	-0.038
	S.D.	0.145	0.136	0.145	0.136	0.143	0.143	0.158
	Skewness	0.033	0.005	0.034	0.004	0.002	0.002	0.131
	Kurtosis	4.303	3.262	4.298	3.266	3.326	3.330	4.597

Table 2. Estimated bias, standard deviation, skewness and kurtosis of the four estimators of the MA(1) parameter when θ = 0.4.

Design Matrix	Statistic	Estimator						
		MML						Max. Likeli. Est.
		ML_1	CML_1	ML_2	CML_2	AML_1	ML_2	
X1 n = 30	Bias	0.046	0.000	0.050	0.004	-0.028	-0.027	-0.025
	S.D.	0.302	0.186	0.308	0.195	0.190	0.193	0.274
	Skewness	0.023	0.711	0.012	0.703	0.687	0.703	0.368
	Kurtosis	4.111	5.292	4.124	5.034	4.165	4.226	7.082
X1 n = 60	Bias	0.015	0.016	0.015	0.016	0.001	0.001	0.015
	S.D.	0.161	0.140	0.162	0.140	0.130	0.130	0.170
	Skewness	1.090	1.092	1.099	1.068	0.024	0.023	0.483
	Kurtosis	6.080	6.756	6.065	6.696	3.205	3.208	5.240
X2 n = 30	Bias	0.056	0.165	0.053	0.159	-0.017	-0.019	-0.066
	S.D.	0.292	0.272	0.296	0.273	0.181	0.183	0.332
	Skewness	0.108	0.304	0.052	0.316	0.704	0.706	1.199
	Kurtosis	4.005	2.132	4.346	2.190	4.841	4.850	7.341
X2 n = 60	Bias	0.008	0.045	0.008	0.045	0.000	-0.000	-0.023
	S.D.	0.147	0.139	0.146	0.139	0.124	0.125	0.145
	Skewness	0.145	1.118	0.112	1.110	0.024	0.024	0.021
	Kurtosis	9.737	6.420	9.734	6.406	3.198	3.206	4.548
X3 n = 30	Bias	0.043	0.071	0.044	0.071	-0.012	-0.012	0.011
	S.D.	0.260	0.251	0.262	0.252	0.173	0.174	0.255
	Skewness	0.280	0.366	0.272	0.355	0.491	0.487	0.108
	Kurtosis	3.950	3.821	3.927	3.792	4.213	4.221	5.966
X3 n = 60	Bias	0.008	0.019	0.008	0.019	0.000	0.000	0.001
	S.D.	0.138	0.136	0.138	0.136	0.123	0.123	0.136
	Skewness	0.485	0.597	0.483	0.593	0.011	0.010	0.052
	Kurtosis	5.725	5.838	5.715	5.827	3.087	3.090	4.173
X4 n = 30	Bias	0.043	-0.004	0.046	0.001	-0.019	-0.017	0.088
	S.D.	0.272	0.230	0.279	0.235	0.180	0.181	0.300
	Skewness	0.141	0.266	0.063	0.253	0.656	0.643	0.039
	Kurtosis	3.970	5.079	4.347	4.901	4.279	4.268	3.333
X4 n = 60	Bias	0.013	0.003	0.013	0.004	-0.000	0.000	0.024
	S.D.	0.158	0.152	0.158	0.152	0.128	0.128	0.157
	Skewness	1.160	1.285	1.129	1.270	0.011	0.011	0.605
	Kurtosis	6.184	6.782	6.153	6.743	3.032	3.034	5.227
X5 n = 30	Bias	0.048	0.106	0.047	0.104	-0.014	-0.014	-0.025
	S.D.	0.269	0.249	0.270	0.251	0.175	0.177	0.274
	Skewness	0.286	0.370	0.275	0.359	0.539	0.537	0.368
	Kurtosis	3.713	3.602	3.711	3.579	4.405	4.411	7.082
X5 n = 60	Bias	0.007	0.031	0.007	0.031	-0.000	-0.000	-0.011
	S.D.	0.140	0.136	0.141	0.136	0.124	0.124	0.139
	Skewness	0.490	0.699	0.486	0.657	0.019	0.019	0.005
	Kurtosis	5.813	5.913	5.804	5.850	3.166	3.171	4.038

Table 3. Estimated bias, standard deviation, skewness and kurtosis of the four estimators of the MA(1) parameter when θ = -0.4.

Design Matrix	Statistic	Estimator						
		MML						Max. Likeli. Est.
		ML_1	CML_1	ML_2	CML_2	AML_1	AML_2	
X1 n = 30	Bias S.D. Skewness Kurtosis	-0.027 0.299 0.057 3.493	0.234 0.163 0.051 4.892	-0.040 0.313 0.007 3.757	0.220 0.168 0.023 5.047	0.047 0.197 0.638 3.806	0.041 0.199 0.645 3.815	-0.307 0.323 0.648 3.464
X1 n = 60	Bias S.D. Skewnes Kurtosis	-0.011 0.179 0.425 6.852	0.094 0.134 0.030 9.365	-0.013 0.178 0.684 5.449	0.090 0.136 0.003 9.467	0.006 0.141 0.072 3.779	0.005 0.142 0.070 3.777	-0.146 0.243 0.444 0.674
X2 n = 30	Bias S.D. Skewness Kurtosis	-0.022 0.289 0.061 3.905	0.269 0.203 0.212 8.599	-0.036 0.297 0.058 3.675	0.249 0.208 0.043 8.328	0.047 0.189 0.536 3.730	0.040 0.191 0.538 3.725	-0.397 0.315 1.261 2.761
X2 n = 60	Bias S.D. Skewnes Kurtosis	-0.010 0.170 0.704 5.147	0.106 0.135 0.060 5.337	-0.012 0.171 0.728 5.129	0.102 0.136 0.121 5.683	0.005 0.139 0.022 2.968	0.003 0.139 0.020 2.964	-0.159 0.237 0.487 2.662
X3 n = 30	Bias S.D. Skewness Kurtosis	-0.026 0.266 0.135 4.338	0.078 0.217 0.027 6.298	-0.032 0.267 0.191 3.942	0.070 0.220 0.037 6.146	0.031 0.177 0.430 3.799	0.028 0.178 0.424 3.794	-0.129 0.291 0.097 2.445
X3 n = 60	Bias S.D. Skewnes Kurtosis	-0.004 0.148 0.554 5.612	0.030 0.138 0.279 5.546	-0.004 0.148 0.574 5.639	0.028 0.139 0.307 5.619	0.006 0.127 0.025 3.068	0.005 0.128 0.025 3.068	-0.042 0.156 0.590 5.214
X4 n = 30	Bias S.D. Skewness Kurtosi	-0.047 0.290 0.248 3.121	-0.001 0.256 0.484 4.180	-0.050 0.292 0.232 3.076	-0.005 0.261 0.348 4.365	0.032 0.180 0.416 3.570	0.031 0.181 0.406 3.560	-0.085 0.312 0.088 2.791
X4 n = 60	Bias S.D. Skewnes Kurtosis	-0.007 0.158 0.933 5.958	0.002 0.149 0.871 6.386	-0.008 0.158 0.925 5.936	0.001 0.150 0.890 6.390	0.005 0.129 0.031 3.085	0.005 0.129 0.030 3.085	-0.019 0.159 0.553 5.173
X5 n = 30	Bias S.D. Skewness Kurtosis	-0.018 0.274 0.061 4.367	0.186 0.191 0.060 6.776	-0.029 0.278 0.100 3.935	0.172 0.198 0.003 6.844	0.041 0.183 0.485 3.734	0.035 0.184 0.482 3.725	-0.264 0.326 0.083 1.767
X5 n = 60	Bias S.D. Skewnes Kurtosis	-0.008 0.161 0.732 5.392	0.065 0.140 0.334 6.003	-0.010 0.162 0.757 5.397	0.062 0.140 0.324 5.925	0.005 0.133 0.018 2.967	0.004 0.133 0.017 2.967	-0.097 0.201 0.909 3.991

Table 4. Estimated loss of seven different estimators of the MA(1) parameter.

Design Matrix	n	θ	Estimators						
			MML						Max. Likeli. Est.
			ML_1	CML_1	ML_2	CML_2	AML_1	AML_2	
X1	30	0	0.280	0.144	0.213	0.221	0.086	0.083	0.306
X1	60	0	0.105	0.146	0.105	0.145	0.077	0.077	0.318
X2	30	0	0.197	0.482	0.196	0.464	0.085	0.082	0.318
X2	60	0	0.103	0.126	0.104	0.124	0.066	0.066	0.302
X3	30	0	0.216	0.212	0.223	0.219	0.088	0.088	0.293
X3	60	0	0.066	0.084	0.066	0.083	0.065	0.065	0.149
X4	30	0	0.265	0.225	0.262	0.248	0.088	0.089	0.334
X4	60	0	0.073	0.067	0.072	0.068	0.070	0.070	0.080
X5	30	0	0.201	0.297	0.199	0.302	0.085	0.083	0.408
X5	60	0	0.103	0.101	0.103	0.202	0.063	0.063	0.165
X1	30	0.4	0.191	0.226	0.195	0.223	0.212	0.215	0.309
X1	60	0.4	0.304	0.324	0.305	0.319	0.055	0.055	0.208
X2	30	0.4	0.203	0.322	0.208	0.315	0.224	0.227	0.471
X2	60	0.4	0.323	0.343	0.318	0.341	0.052	0.052	0.131
X3	30	0.4	0.196	0.226	0.196	0.224	0.170	0.170	0.209
X3	60	0.4	0.209	0.237	0.208	0.236	0.046	0.046	0.096
X4	30	0.4	0.185	0.187	0.196	0.178	0.199	0.196	0.205
X4	60	0.4	0.313	0.337	0.308	0.335	0.045	0.045	0.226
X5	30	0.4	0.196	0.253	0.194	0.249	0.184	0.185	0.309
X5	60	0.4	0.213	0.262	0.212	0.255	0.050	0.050	0.097
X1	30	-0.4	0.152	0.365	0.173	0.355	0.214	0.209	0.504
X1	60	-0.4	0.261	0.378	0.243	0.375	0.091	0.089	0.288
X2	30	-0.4	0.159	0.567	0.167	0.521	0.197	0.190	0.651
X2	60	-0.4	0.224	0.245	0.229	0.261	0.057	0.054	0.307
X3	30	-0.4	0.180	0.275	0.178	0.265	0.167	0.164	0.258
X3	60	-0.4	0.212	0.201	0.216	0.206	0.054	0.053	0.241
X4	30	-0.4	0.175	0.184	0.176	0.182	0.160	0.157	0.207
X4	60	-0.4	0.274	0.274	0.272	0.276	0.055	0.055	0.215
X5	30	-0.4	0.167	0.396	0.168	0.381	0.183	0.177	0.428
X5	60	-0.4	0.232	0.260	0.237	0.253	0.053	0.052	0.302