ISSN 1440-771X
ISBN 0 7326 1042 7

MONASH UNIVERSITY

AUSTRALIA

Modified Likelihood and Related Methods for Handling Nuisance Parameters in the Linear Regression Model

Mizan R. Laskar and Maxwell L. King

Working Paper 5/98
June 1998

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

MODIFIED LIKELIHOOD AND RELATED METHODS FOR HANDLING NUISANCE PARAMETERS IN THE LINEAR REGRESSION MODEL

Mizan R. Laskar and Maxwell L. King

Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3168, Australia

Summary

In this paper, different approaches to dealing with nuisance parameters in likelihood based inference are presented and illustrated by reference to the linear regression model with nonspherical errors. The estimator of the error variance using each of the approaches is also derived for the linear regression model with spherical errors. We observe that many of these estimators are unbiased. A theoretical comparison of the likelihood functions is reported and we note that some of them are equivalent. Empirical evidence in the literature indicates that estimators based on the conditional profile likelihood and tests based on the marginal likelihood have better small sample properties compared to those based on other likelihood and message length functions.

Key words: Linear regression errors, parameter orthogonality, marginal likelihood, modified profile likelihood, message length function.

1 Introduction

Satisfactory statistical analysis of non-experimental data, is an important problem in statistics and especially econometrics. Often in such cases, statistical models involve a large number of influences, most of which are not of immediate interest. This means that such models contain two kinds of parameters, those of interest and those not of immediate interest that are known as nuisance parameters. Their presence causes unexpected complications in statistical inference.

A fairly standard procedure for making inferences about any parameter of interest is to replace the nuisance parameters by their respective maximum likelihood (ML) estimators. In such situations, estimators and tests can perform poorly in small samples (Bewley, 1986, Cox and Reid, 1987, King, 1987, King and McAleer, 1987, Moulton and Randolph, 1989 and Chesher and Austin, 1991). An early example of such a problem was drawn to the attention of the statistical profession by Cochrane and Orcutt (1949). They showed that the von Neumann ratio, designed to test for autocorrelation in an observed time series, was biased towards accepting randomness when applied to ordinary least squares (OLS) residuals from a linear regression. In this example, the regression coefficients are nuisance parameters and in order to test the regression errors, these coefficients are replaced by their OLS estimators. Cochrane and Orcutt's timely warning gave raise to the familiar Durbin-Watson test. Earlier, Neyman and Scott (1948) warned that nuisance parameters can seriously compromise likelihood based inference. In this connection, King (1996) observed that when nuisance parameters are present, statistical theory is generally less helpful in suggesting reliable diagnostic tests. Also, Cordus (1986) noted that the presence of nuisance parameters causes a shift in the estimated mean of the null distribution of the likelihood ratio test. The question then arises: how to tackle the problem of nuisance parameters in order to improve estimators and tests?

There is a vast amount of literature on the satisfactory handling of nuisance parameters, the application of which can improve likelihood based estimators and test procedures. Most of this work has focused on the modification of the likelihood function and the profile (or concentrated) likelihood function. In this context, Kalbfleisch and Sprott (1970) derived the marginal likelihood function and conditional likelihood function as a method of eliminating nuisance parameters.

Subsequently, Ara (1995), Ara and King (1993, 1995) and Rahman and King (1998) used the marginal likelihood to construct different tests for linear regression disturbance parameters and observed a significant improvement in small sample properties over those of conventional tests. A related approach known as residual (or restricted) maximum likelihood (REML) (Patterson and Thompson, 1971) has gained considerable importance in the context of estimating variance components in the linear regression model. As an alternative approach to handling nuisance parameters, Cox and Reid (1987) initiated the idea of the conditional profile likelihood (CPL). In addition, Barndorff-Nielsen (1983) proposed the modified profile likelihood (MPL) function and McCullagh and Tibshirani (1990) suggested a slightly different way of handling nuisance parameters based on a simple adjustment to the profile score statistic. More recently, Macaskill (1993) extended this work to multiparameter non-linear regression problems. A similar approach based on the expected log likelihood was proposed by Conniffe (1987) and involves equating the score vector to its expected value and solving for the unknown parameters. A slightly different approach based on singular value decomposition for the likelihood function was proposed by Hinde and Aitkin (1987). On the other hand, Wallace and Freeman (1987) introduced the idea of the minimum message length (MML) estimator with a Bayesian viewpoint as an alternative method of estimation for the parameter of interest. Extending this research, Wallace and Freeman (1992) and Wallace and Dowe (1994) applied the MML estimation method to different problems and observed that it gives improved estimates compared to the ML estimator.

As can be seen from this brief survey of the literature, there is a vast array of suggestions for handling nuisance parameters. However, these various approaches are not equally efficient for statistical problems where estimation and diagnostic testing are of main interest. The purpose of this article is to examine each of the approaches with an eye to their applicability in the regression model. In particular, their application is investigated in the context of inference involving parameters of the error process in the general linear model.

Following the above introduction, this paper is divided into five sections. Section 2 deals with the derivation of the functional form for each of the approaches. Estimators of the error variance for the simple linear regression model that were

obtained using each of the approaches are presented in section 3. Section 4 deals with theoretical comparisons as well as a review of emperical comparisons of different methods for estimation, testing, model selection and forecasting. Finally, some concluding remarks are made in section 5.

2 Methods That Deal With Nuisance Parameters

Consider the linear regression model with non-spherical disturbances

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

where y is $n\times 1$, X is $n\times k$, nonstochastic and of rank $k < n$, β is a $k\times 1$ vector, $\Omega(\theta)$ is a symmetric matrix and θ is a $p\times 1$ vector. This model generalizes a wide range of disturbance processes of the linear regression model of particular interest to statisticians and econometricians. 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 models including those that result from random regression coefficients. The likelihood and log likelihood for this model (excluding constants) are respectively

$$L(y;\theta,\sigma^2,\beta) \propto \sigma^{-n}\left|\Omega(\theta)\right|^{-\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(y - X\beta)'\Omega(\theta)^{-1}(y - X\beta)\}, \tag{2}$$

$$l(y;\theta,\sigma^2,\beta) \propto -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\log\left|\Omega(\theta)\right| - \frac{1}{2\sigma^2}(y - X\beta)'\Omega(\theta)^{-1}(y - X\beta) \tag{3}$$

and the log profile (or concentrated) likelihood is

$$l_p(y;\theta) \propto -\frac{n}{2}\log\hat{\sigma}_\theta^2 - \frac{1}{2}\log\left|\Omega(\theta)\right| \tag{4}$$

where $\hat{\sigma}_\theta^2 = (y - X\hat{\beta}_\theta)'\Omega(\theta)^{-1}(y - X\hat{\beta}_\theta)/n$ and $\hat{\beta}_\theta = (X'\Omega(\theta)^{-1}X)^{-1}X'\Omega(\theta)^{-1}y$.
In the subsequent sub-sections, different likelihoods are illustrated by their application to model (1) with a view to making inferences about θ.

2.1 Marginal Likelihood

As a useful method for eliminating nuisance parameters, the concept of the marginal likelihood was first introduced by Fraser (1967) in the structural inference context, and further developed by Kalbfleisch and Sprott (1970) in the classical framework. The key idea is to transform y to another random vector, a subvector of

which has a likelihood (marginal likelihood) that only involves the parameters of interest and the remainder of which contains no information about those parameters. Tunnicliffe Wilson (1989) derived the marginal likelihood for θ in (1) as

$$L_m(y;\theta) = \left|\Omega(\theta)\right|^{-\frac{1}{2}} \left|X'\Omega(\theta)^{-1}X\right|^{-\frac{1}{2}} \hat{s}^{-\frac{m}{2}} \tag{5}$$

where $\hat{s} = (y - X\hat{\beta}_\theta)'\Omega(\theta)^{-1}(y - X\hat{\beta}_\theta)$ and $m = n - k$.

Ara and King (1993) developed marginal likelihood based likelihood ratio (LR), Lagrange multiplier (LM), Wald and King and Wu's (1997) asymptotically locally most mean powerful (ALMMP) tests for the covariance matrices of regression disturbances. They pointed out that the problem of testing different θ values is invariant under the transformation $y \rightarrow \eta_0 y + X\eta$ where η_0 is a positive scalar and η is a $k\times 1$ vector. They also demonstrated that these tests can be constructed by treating the maximal invariant statistic, $w = L'z / (z'LL'z)^{\frac{1}{2}}$ as the observed data where L is an $n\times m$ matrix such that $L'L = I_m$ and $LL' = I_n - X(X'X)^{-1}X'$, and z is the OLS residual vector from (1). The density function of the maximal invariant statistic is

$$f(w;\theta) = \frac{1}{2}\,\Gamma(m/2)\pi^{-\frac{m}{2}}\,\left|L'\Omega(\theta)L\right|^{-\frac{1}{2}}\hat{q}^{-\frac{m}{2}} \tag{6}$$

where $\hat{q} = w'(L'\Omega(\theta)L)^{-1}w = \hat{u}'\Omega(\theta)^{-1}\hat{u} / z'z = \hat{s} / z'z$, $\hat{u}$ is the generalized least squares (GLS) residual vector assuming the covariance matrix $\sigma^2\Omega(\theta)$ and

$$\left|L'\Omega(\theta)L\right| = \left|X'X\right|^{-1}\left|\Omega(\theta)\right|\left|X'\Omega(\theta)^{-1}X\right|$$

(Verbyla, 1990 and Ara and King, 1993). The theory of invariance implies that all invariant tests can be constructed by treating w as observed data.

Constructing a marginal likelihood involves dividing the information in the data into two parts by means of ancillary statistics, one of which contains θ only and the other being uninformative about θ. The marginal likelihood function and the likelihood of the maximal invariant statistic are equivalent in this case because the ratio of (5) and (6) is independent of θ. The marginal likelihood function for θ is therefore a likelihood function and enjoys the properties of a likelihood. A draw-back is that the marginal likelihood generally cannot be defined for nonlinear regression, because the required transformation typically does not exist. Levenbach (1972)

introduced the marginal likelihood for the parameters of a Gaussian autoregressive process from the marginal distribution of a vector of standardized residuals resulting from the conditional techniques of structural inference. Bellhouse (1978) demonstrated the application of the marginal likelihood approach to ARMA models, lagged dependent variable regression models and polynomial distributed lagged regression models and discussed the treatment of nuisance parameters for such models.

2.2 Residual Likelihood

Patterson and Thompson (1971) introduced the idea of REML estimation in the case of unbalanced incomplete block designs. It was subsequently generalized by Thompson (1973), while Harville (1974) showed the residual likelihood is equivalent to the marginal likelihood for all the regression disturbance parameters; i.e. θ and σ^2. In a similar context, Verbyla (1990) presented an alternative derivation of the residual likelihood for regression disturbance parameters which is based on the log likelihood

$$l_r(y;\theta,\sigma^2) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|L'\Omega(\theta)L| - \frac{\hat{s}}{2\sigma^2}. \tag{7}$$

2.3 Modified Profile Likelihood

Barndorff-Nielsen (1983) is responsible for the basic idea of the MPL where the profile likelihood is adjusted by two factors. The MPL function, denoted by $L_{mp}(y;\theta)$, is given by

$$L_{mp}(y;\theta) = \left|\frac{\partial\hat{\gamma}^{\dagger}}{\partial\hat{\gamma}_{\theta}^{\dagger}}\right| \left\{\left|-\frac{\partial^2 l(y;\theta,\sigma^2,\beta)}{\partial\gamma^{\dagger}\partial\gamma^{\dagger}{}'}\right|_{\hat{\gamma}_{\theta}^{\dagger}}\right\}^{-\frac{1}{2}} L_p(y;\theta) \tag{8}$$

where $\gamma^{\dagger} = (\sigma^2,\beta')'$, $\gamma = (\theta',\sigma^2,\beta')'$, $\frac{\partial\hat{\gamma}^{\dagger}}{\partial\hat{\gamma}_{\theta}^{\dagger}}$ is the matrix of partial derivatives of $\hat{\gamma}^{\dagger}$ with respect to $\hat{\gamma}_{\theta}^{\dagger}$, $\hat{\gamma}^{\dagger}$ is the overall ML estimator of $\gamma^{\dagger}$, $\hat{\gamma}_{\theta}^{\dagger}$ is the ML estimator of $\gamma^{\dagger}$ for fixed θ and $L_p(y;\theta)$ is the profile likelihood for θ. The modifying factor

$\left\{\left|-\frac{\partial^2 l(y;\theta,\sigma^2,\beta)}{\partial\gamma^{\dagger}\partial\gamma^{\dagger}{}'}\right|_{\hat{\gamma}^{\dagger}_{\theta}}\right\}^{-\frac{1}{2}}$ corresponds to the variance stabilization transformation of the parameter θ and $\left|\frac{\partial\hat{\gamma}^{\dagger}}{\partial\hat{\gamma}^{\dagger}_{\theta}}\right|$ is a correction for parameterization that ensures invariance under reparameterization. The difficulty with applying this formula, as mentioned by Cox and Reid (1987), is that it requires conditioning on an appropriate ancillary statistic. The log of (8) is

$$l_{mp}(y;\theta) = \log\left|\frac{\partial\hat{\gamma}^{\dagger}}{\partial\hat{\gamma}^{\dagger}_{\theta}}\right| - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\theta,\sigma^2,\beta)}{\partial\gamma^{\dagger}\partial\gamma^{\dagger}{}'}\right|_{\hat{\gamma}^{\dagger}_{\theta}} + l_p(y;\theta).$$

2.4 Conditional Profile Likelihood

This method has been given different names by different authors. Cox and Reid (1987) called it the CPL, Simonoff and Tasi (1994) and Ferguson et al. (1991) named it the MPL, Barndorff-Nielsen and McCullagh (1993) denoted it the adjusted profile likelihood, Mukerjee (1993) called it the conditional likelihood and Fraser and Reid (1989) named it the approximate conditional likelihood. To avoid this confusion, we will refer to it as the CPL following Cox and Reid (1987), who introduced this approach.

Cox and Reid (1987) pointed out that inferences based on the profile likelihood are inefficient due to lack of orthogonality. They explored a modification to the profile likelihood, in which the nuisance parameter is reparameterized to be orthogonal to the parameter of interest. This approach works in two steps. First, the nuisance parameter is made orthogonal to the parameter of interest which can be achieved by solving a differential equation. Then, a correction is made to the profile likelihood function of the transformed model. Let us first examine the orthogonality issue for the parameters of the model (1).

2.4.1 Orthogonality

The key feature in the application of the CPL to model (1) is the derivation of transformed parameters whose asymptotic covariances are zero. Two parameters, η

and ζ say, are orthogonal if the (η, ζ) element of the information matrix is zero. The information matrix for the vector $\gamma = (\theta', \sigma^2, \beta')'$ is given by

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

where the $(i,j)^{\text{th}}$ element of $A(\theta)$ is $\frac{1}{2}tr\left[-\frac{\partial\Omega(\theta)}{\partial\theta_i}\frac{\partial\Omega(\theta)^{-1}}{\partial\theta_j}\right]$ and the i^{th} element of $B(\theta)$ is $\frac{1}{2\sigma^2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta_i}\right]$. It is observed from the above information matrix that the parameters (θ, σ^2) and β are orthogonal but θ and σ^2 are not orthogonal. The first step in constructing the CPL is to make an orthogonal transformation $\gamma = (\theta', \sigma^2, \beta')' \rightarrow \gamma_m = (\theta', \delta, \beta')'$ so that the asymptotic covariance of the ML estimators of θ and δ are zero. In this context, Cox and Reid (1987) mentioned that the parameter θ should be scalar; otherwise, global orthogonality cannot always be obtained. In this approach, we assume $p = 1$, so that θ is a scalar. We want a transformation from (θ, σ^2) to (ϕ_1, ϕ_2) keeping $\theta = \phi_1$ fixed and adjusting $\phi_2 = \phi_2(\theta, \sigma^2)$ so that θ and ϕ_2 are orthogonal. Huzurbazar (1950) noted that the transformation is established by solving

$$i^*_{\theta,\phi_2} = i_{\theta,\sigma^2} + i_{\sigma^2,\sigma^2}\frac{\partial\sigma^2}{\partial\theta} = 0 \qquad (9)$$

where $i^*_{\theta,\phi_2} = E\left(-\frac{\partial^2 l}{\partial\theta\partial\phi_2}\right)$ is the information measure calculated in the (θ, ϕ_2) parameterization, $i_{\theta,\sigma^2} = E\left(-\frac{\partial^2 l}{\partial\theta\partial\sigma^2}\right)$ and $i_{\sigma^2,\sigma^2} = E\left(-\frac{\partial^2 l}{\partial\sigma^2\partial\sigma^2}\right)$ are the information measures calculated for the (θ, σ^2) and (σ^2, σ^2) parameters, respectively. For the above information matrix, equation (9) becomes

$$\frac{n}{2\sigma^4}\frac{\partial\sigma^2}{\partial\theta} = -\frac{1}{2\sigma^2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right].$$

There is a degree of arbitrariness in the solution of this differential equation. One possible solution is

$$\sigma^2 = \frac{\delta}{|\Omega(\theta)|^{\frac{1}{n}}} \text{ or } \delta = \sigma^2 |\Omega(\theta)|^{\frac{1}{n}}.$$

Laskar and King (1998) mentioned that for model (1), this also works when θ is a vector. Mukerjee (1993) derived the LR test based on the CPL in a general multiparameter set-up. Cox and Reid (1987, p.3) discussed a similar type of reparameterization.

2.4.2 Derivation of the Conditional Profile Likelihood

Returning to the case of $p > 1$, the log likelihood (3) after the orthogonal transformation to γ_m has the form

$$l(y;\gamma_m) = -\frac{n}{2}\log\delta - \frac{1}{2\delta}(y_\theta - X_\theta\beta)'(y_\theta - X_\theta\beta) \tag{10}$$

where $y_\theta = G(\theta)^{-\frac{1}{2}}y$, $X_\theta = G(\theta)^{-\frac{1}{2}}X$, and $G(\theta)$ is an $n\times n$ matrix comprised of $\Omega(\theta)$ with each element divided by $|\Omega(\theta)|^{\frac{1}{n}}$. The log CPL for θ is

$$l_{cp}(y;\theta) = l_p^*(y;\theta) - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\gamma_m)}{\partial\gamma_m^*\partial\gamma_m^{*\prime}}\right|_{\hat{\gamma}_m^*}$$

where $\gamma_m^* = (\delta,\beta')'$, $l_p^*(y;\theta) = -\frac{n}{2}\log\hat{\delta}_\theta - \frac{n}{2}$, $\hat{\delta}_\theta = [(y_\theta - X_\theta\hat{\beta}_\theta^*)'\times (y_\theta - X_\theta\hat{\beta}_\theta^*)]/n$, $\hat{\beta}_\theta^* = (X_\theta'X_\theta)^{-1}X_\theta'y_\theta$ and $\hat{\gamma}_m^*$ is the ML estimator of γ_m^* for fixed θ. After some algebraic manipulation and ignoring constant terms, we get

$$l_{cp}(y;\theta) = -\frac{n}{2}\log\hat{\delta}_\theta - \frac{1}{2}\log\left\{\frac{n}{2\hat{\delta}_\theta^{\,k+2}}|X_\theta'X_\theta|\right\}$$

$$= \log\hat{\delta}^{\frac{k+2-n}{2}} + \log|X_\theta'X_\theta|^{-\frac{1}{2}} + \log\left(\frac{n}{2}\right)^{-\frac{1}{2}}$$

or

$$l_{cp}(y;\theta) = -\frac{m-2}{2}\log\hat{\delta}_\theta - \frac{1}{2}\log|X_\theta'X_\theta|.$$

Therefore the CPL without its constant terms is

$$L_{cp}(y;\theta) = |X_\theta'X_\theta|^{-\frac{1}{2}}\hat{\delta}_\theta^{-\frac{m-2}{2}}. \tag{11}$$

The effect of the second term of (11) is to penalize those values of θ that give relatively high information about σ^2.

A similar derivation for the simple linear regression model with heteroscedastic error variances was introduced by Simonoff and Tasi (1994). Cox (1988) mentioned that for simple exponential family problems, this procedure performs well. It is very close to the REML procedure for estimating variance components. On the other hand, Ferguson (1992) observed that this method has the disadvantage of the non-uniqueness of orthogonal parameterization. If θ and δ are orthogonal parameters, it is also true that $h(\delta)$ and θ are orthogonal for any continuous function h.

Ferguson et al. (1991) discussed the properties of the score equation derived from the CPL using a stochastic asymptotic expansion. They considered the relationship between the derivative of the score function and its variance and observed that the CPL is not a true likelihood function, so it does not have all the properties of a likelihood function. For example, this likelihood does not have the property that its second order derivative has a negative mean value equal to the variance of the score statistic. They also pointed out that (11) is not invariant under reparameterizations of the nuisance parameter σ^2 and under non-linear transformations.

2.5 Conditional profile restricted log-likelihood

Laskar and King (1998) identified that expression (7) involves the nuisance parameter σ^2. Its presence may cause problems in small samples for estimators and tests of elements of θ based on $l_r(y;\theta,\sigma^2)$. They eliminated the effect of σ^2 from (7) by combining Cox and Reid's (1987) CPL method outlined in section 2.4 and called it the conditional profile restricted likelihood (CPRL), which in log form for θ is

$$\bar{l}^*_{cpr}(y;\theta) = -\frac{m-2}{2m}\log|\Omega(\theta)| - \frac{m-2}{2m}\log|X'\Omega(\theta)^{-1}X| - \frac{m-2}{2}\log(\hat{u}'\Omega(\theta)^{-1}\hat{u}). \qquad (12)$$

2.6 Canonical Likelihood

Hinde and Aitkin (1987) proposed a different approach to handling nuisance parameters based on a singular value decomposition of the likelihood function.

Consider the likelihood function $L(y;\theta,\sigma^2,\beta)$ in (2). This likelihood function can be denoted by $L(\theta,\theta^*)$, where $\theta^* = (\sigma^2,\beta')'$. The idea behind the canonical likelihood, is to divide the likelihood function $L(\theta,\theta^*)$ into $L_1(\theta)$ and $L_2(\theta^*)$ such that

$$\int L(\theta,\theta^*)L_2(\theta^*)d\theta^* = \lambda L_1(\theta) \tag{13}$$

$$\int L(\theta,\theta^*)L_1(\theta)d\theta = \lambda L_2(\theta^*) \tag{14}$$

which minimize

$$\iint \{L(\theta,\theta^*) - L_1(\theta)L_2(\theta^*)\}^2 d\theta d\theta^*$$

where λ^2 is the principal eigenvalue of L a $p\times q$ matrix defined as follows. For distinct parametric points θ_i,θ_j^*, $i = 1, 2, ..., p$, $j = 1, 2, ..., q$, the likelihood function $L(\theta,\theta^*)$ can be written as a $p\times q$ matrix L with elements $a_{ij} = L(\theta_i,\theta_j^*)$. $L_1(\theta)$ and $L_2(\theta^*)$ are given by the principal left and right eigenvectors of the likelihood matrix L. Hinde and Aitkin (1987) show that these two equations can be written as homogeneous Fredholm equations of the second kind. Substituting equation (13) into equation (14) gives

$$\lambda^2 L_1(\theta) = \int K(\theta,\phi)L_1(\phi)d\phi$$

where the symmetric kernel function is given by

$$K(\theta,\phi) = \int L(\theta,\theta^*)L(\phi,\theta^*)d\theta^*.$$

In some cases, analytical integration over θ^* is possible, giving a q dimensional kernel function, but for most cases no analytical solution exists and one needs to resort to numerical solutions. The authors argued that canonical likelihoods can be found from any two-parameter model, though marginal and conditional likelihoods may not be found. They demonstrated the application of this method in several two parameter models and explained the possibility of its application in multi-parameter models, where integration is needed for the k dimensional kernel function. The authors did not mention its use in inference. We are therefore not sure whether its use can improve estimators and tests. The application of this approach is limited in practice, especially in econometric models containing large numbers of parameters, because for such models, no analytical solution is possible for L_1 and L_2.

2.7 Expected Log-Likelihood Approach

The usual ML estimator is obtained by setting the score equal to zero and solving for θ. An alternative approach, called the expected maximum likelihood estimator (EMLE), was introduced by Conniffe (1987) and involves equating the score function to its expected value. This approach is based on the fact that the true values of the parameters maximize the expected log likelihood rather than the actual log likelihood. So, determining the true value of the parameters is the same as determining the maximum of the expected log likelihood. Exact algebraic expressions for the resultant estimators may not always be possible, but a generally applicable approximate procedure is given by Conniffe (1988, 1990a). If $\hat{\beta}$ and $\hat{\sigma}^2$ are the available estimators of β and σ^2 respectively, an estimator of θ can be obtained by setting

$$\left[\frac{\partial}{\partial\theta_i} l(y;\theta,\sigma^2,\beta)\right]_{\hat{\gamma}_\theta^t}, \; i = 1, 2, ..., p, \tag{15}$$

equal to their expectations, or to an approximation of their expectations, and solving for θ. We have

$$\left.\frac{\partial l(y;\theta,\sigma^2,\beta)}{\partial\theta_i}\right|_{\hat{\gamma}_\theta^t} = -\frac{1}{2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta_i}\right] - \frac{n}{2}\left[\hat{u}'\frac{\partial\Omega(\theta)^{-1}}{\partial\theta_i}\hat{u} / \hat{u}'\Omega(\theta)^{-1}\hat{u}\right]. \tag{16}$$

Using the results of Ara and King (1993) and Mahmood and King (1997)

$$E\left[\left.\frac{\partial l(y;\theta,\sigma^2,\beta)}{\partial\theta_i}\right|_{\hat{\gamma}_\theta^t}\right] = -\frac{1}{2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta_i}\right] + \frac{n}{2m}tr\left[P_\theta\frac{\partial\Omega(\theta)}{\partial\theta_i}\right] \tag{17}$$

where $P_\theta = \Omega(\theta)^{-1} - \Omega(\theta)^{-1}X(X'\Omega(\theta)^{-1}X)^{-1}X'\Omega(\theta)^{-1}$. Equating (16) with (17), we have

$$\frac{n}{2m}tr\left[P_\theta\frac{\partial\Omega(\theta)}{\partial\theta_i}\right] = -\frac{n}{2}\left[\hat{u}'\frac{\partial\Omega(\theta)^{-1}}{\partial\theta_i}\hat{u} / \hat{u}'\Omega(\theta)^{-1}\hat{u}\right]. \tag{18}$$

The estimate of θ is the iterative solution for θ from (18). Conniffe (1990a) mentioned that for a single test parameter, the estimated score test is based on the difference between the estimated score and its expectation, but in some cases it can lead to expressions which have no exact algebraic solutions. Also, Conniffe (1990b) observed that the first-order asymptotic properties of the estimated score test and LM

test coincide. However, the small sample and higher-order asymptotic properties of these tests may differ.

2.8 Adjusted Profile Likelihood

This is another approach to handling nuisance parameters, pioneered by McCullagh and Tibshirani (1990). In this case, an adjustment is made to the profile likelihood score function to make the mean of the score function equal to zero and the variance of the score function equal to its negative expected derivative matrix. Two properties of the profile likelihood score statistic are recovered by this approach, namely, an adjustment that makes the mean of the score statistic zero again and its variance equal to minus the expected derivative of the score. Macaskill (1993) extended this approach to multiparameter non-linear regression models.

The profile likelihood $l_p(y;\theta)$ for model (1) is given by (4). The i^{th} element of the score vector, $S(\theta)$, is given by

$$S_i(\theta) = \frac{\partial}{\partial \theta_i} l_p(y;\theta) = -\frac{1}{2} tr\left[\Omega(\theta)^{-1} \frac{\partial}{\partial \theta_i}\Omega(\theta)\right] - \frac{n}{2}\left[\hat{u}' \frac{\partial \Omega(\theta)^{-1}}{\partial \theta_i} \hat{u} / \hat{u}'\Omega(\theta)^{-1}\hat{u}\right].$$

The key idea is to correct $S(\theta)$ by a $p\times 1$ mean adjustment vector $m(\theta)$ and a $p\times p$ covariance adjustment matrix $W(\theta)$. The required adjusted score function is

$$\tilde{S}(\theta) = W(\theta)\{S(\theta) - m(\theta)\}.$$

The conditions for this adjustment are

$$E_{\theta,\hat{\gamma}_\theta^\dagger}(\tilde{S}(\theta)) = 0$$

and

$$\text{var}_{\theta,\hat{\gamma}_\theta^\dagger}(\tilde{S}(\theta)) = -E_{\theta,\hat{\gamma}_\theta^\dagger}(B(\theta))$$

where the expectations are computed under $(\theta,\hat{\gamma}_\theta^\dagger)$ instead of the true parameter point and the $(i,j)^{\text{th}}$ element of $B(\theta)$ is given by $\frac{\partial}{\partial \theta_j}\tilde{S}_i(\theta)$. Solving the above equations for $m(\theta)$ and $W(\theta)$, we get

$$m(\theta) = E_{\theta,\hat{\gamma}_\theta^\dagger}(S(\theta))$$

and

$$W(\theta) = [\text{var}_{\theta,\hat{\gamma}_\theta^\dagger}\{S(\theta)\}]^{-1}[-E_{\theta,\hat{\gamma}_\theta^\dagger}\{H(\theta)\} + \psi(\theta)]'$$

where the $(i,j)^{th}$ elements of $H(\theta)$ and $\psi(\theta)$ are $\frac{\partial}{\partial\theta_j}S_i(\theta)$ and $\frac{\partial}{\partial\theta_j}m_i(\theta)$, respectively. Finally the adjusted log profile likelihood is given by

$$l_{ap}(y;\theta) = \int^{\theta} \tilde{S}(t)dt. \tag{19}$$

The exponential of (19) is called the adjusted profile likelihood.

McCullagh and Tibshirani (1990) mentioned that sometimes expression (19) can be computed analytically, but in general Monte Carlo simulation is required. They discussed the steps involved in the Monte Carlo simulation of the bootstrap sample to calculate $m(\theta)$ and $W(\theta)$ for each value of θ over a grid of p dimensional space. This likelihood is invariant to reparameterizations. The adjustment of the profile likelihood score function is designed to improve the asymptotic behaviour of likelihood based estimators and tests, but the authors could not provide strong arguments in favour of their desired goal. It does, however, have two appealing features: firstly, the centering of the profile likelihood function, which may improve the consistency of estimators and secondly, the rescaling of the profile likelihood score function, which may improve the second order approximation to its variance and chi-square approximation to the null distribution of LR and Wald statistics.

2.9 Average Likelihood

Aitkin (1991) developed a general likelihood inferential framework for arbitrary model comparison problems, including problems of inference about a single parameter. This likelihood is called the average likelihood and is based on a Bayesian framework. Let $\pi(\theta,\sigma^2,\beta) = \frac{1}{\sigma}$ denote the usual non-informative (improper) prior for (θ,σ^2,β). The average likelihood for fixed θ can be defined as

$$L^A(y;\theta) = \frac{\int L^2(y;\theta,\sigma,\beta)\pi(\theta,\sigma,\beta)d\sigma d\beta}{\int L(y;\theta,\sigma,\beta)\pi(\theta,\sigma,\beta)d\sigma d\beta}.$$

and for θ is given by

$$L^A(y;\theta) = \left|\Omega(\theta)\right|^{-\frac{1}{2}} 2^{-\frac{n}{2}} \frac{\Gamma\{(2n-k-2)/2\}}{\Gamma\{(n-k-2)/2\}} \hat{s}^{-\frac{n}{2}}. \tag{20}$$

Aitkin (1993) mentioned that the average likelihood is a simple penalized form of the profile likelihood. The penalty constant in $L^A(y;\theta)$ does not involve θ. The average likelihood and the profile likelihood are equivalent for the model (1), which is unfortunate, given the aim of this survey.

2.10 Approximate Conditional Profile Likelihood

The major drawback of the CPL is the nonuniqueness of the orthogonal parameterization and the fact that it may not always be possible to find an orthogonal parameterization. These two problems were resolved by Cox and Reid (1993) who derived an approximation to the CPL. This approximation does not require orthogonalization. For model (1), the log ACPL for θ in the scalar case of $p=1$ is given by

$$l_{acp}(y;\theta) = l_p(y;\theta) - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\theta,\sigma^2,\beta)}{\partial\gamma^{\dagger}\partial\gamma^{\dagger}{}'}\right|_{\hat{\gamma}^{\dagger}_{\theta}} + \hat{B}(\theta-\hat{\theta}) \qquad (21)$$

where $\gamma^{\dagger}$ and $\hat{\gamma}^{\dagger}_{\theta}$ are defined in section 2.3, $\hat{\theta}$ is the ML estimator of θ, $\hat{B}$ is $\frac{\partial}{\partial\sigma^2}\{i_{\theta,\sigma^2}.i^{\sigma^2,\sigma^2}\}$ evaluated at $(\hat{\theta},\hat{\sigma}^2)$ and i^{σ^2,σ^2} is the (σ^2,σ^2) element from the inverse of the information matrix. Using the results of section 2.4.1,

$$B = \frac{\partial}{\partial\sigma^2}\left\{\frac{2\sigma^4}{n}\frac{1}{2\sigma^2}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}$$

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

The resulting log ACPL function ignoring the constant term is

$$l_{acp}(y;\theta) = -\frac{n}{2}\log\hat{\sigma}^2_{\theta} - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}\log\left\{\frac{n}{2\hat{\sigma}^{2(k+2)}_{\theta}}\left|X'\Omega^{-1}(\theta)X\right|\right\}$$

$$+\frac{1}{n}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta})$$

or

$$l_{acp}(y;\theta) = -\frac{m-2}{2}\log(\hat{u}'\Omega(\theta)^{-1}\hat{u}) - \frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right|$$

$$+\frac{1}{n}tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]_{\theta=\hat{\theta}}(\theta-\hat{\theta}). \tag{22}$$

2.11 Minimum Message Length

Minimum message length is a Bayesian method which chooses estimators to minimize the length of an encoded form of the data made up of a model and the deviations from that model (residuals). Wallace and Dowe (1993) state that the MML principle is that the best possible conclusion to draw from the data is the theory which maximizes the product of the probability of the data occurring in the light of the theory with the prior probability of that theory.

Let x denote the data and H denote a hypothesis in the form of a model with prior probability Pr(H). The posterior probability becomes

$$\Pr(H|x) = \Pr(H \cup x) / \Pr(x) = \Pr(H)\Pr(x|H) / \Pr(x).$$

We seek an hypothesis or model H on the basis of the knowledge of x and Pr(x) that optimally explains x. This can be viewed as the problem of choosing H to maximize $\Pr(H|x)$ or $\Pr(H)\Pr(x|H)$. We know from the elementary information-theoretic coding that an event, E can be coded by a (binary) message of length(E) = $-\log_2 P(E)$ where P(E) > 0 is the probability of the event E (Wallace and Dowe 1993). Now

$$-\log_2(\Pr(H)\Pr(x|H)) = -\log_2(\Pr(H)) - \log_2(\Pr(x|H)).$$

Maximizing $\Pr(H|x)$ is equivalent to minimizing $-\log_2(\Pr(H)) - \log_2(\Pr(x|H))$, for choosing H. The term $-\log_2(\Pr(H))$ gives the message length of the model and the term $-\log_2(\Pr(x|H))$ gives the message length of the data given the model. Thus we are considering a two part message for describing the data, firstly the model and secondly, the data given this model. Hence the name "minimum message length" (principle) for selecting a model, H to fit observed data, x.

Let $L(x,\mu)$ be the likelihood function for given data x and parameter μ of dimension $p\times 1$, $\pi(\mu)$ be the prior distribution of μ and $F(\mu) = \left|-E\left(\frac{\partial^2 \log L(x;\mu)}{\partial\mu\partial\pi'}\right)\right|$ be the determinant of information matrix. The MML estimate of μ is (Wallace and Freeman, 1987, p. 245) that value of μ which minimize the message length

$$-\log\left(\frac{\pi(\mu)L(x;\mu)}{\sqrt{F(\mu)}}\right)+\frac{p}{2}(1+\log K_p) \qquad (23)$$

where K_p is the p dimensional lattice constant which is independent of parameters, as given by Conway and Sloan (1988, p. 59-61). For example $K_1=\frac{1}{12}$, $K_2=\frac{5}{36\sqrt{3}}$ and $K_3=\frac{19}{36.\sqrt[3]{2}}$. Wallace and Dowe (1994) mentioned, maximizing (23) is equivalent to maximizing the average of the log likelihood function over region of size proportional to $1/\sqrt{F(\mu)}$ while the ML estimator maximizes the likelihood function at a single point. The value of μ which minimizes (23) is the MML estimate of μ with accuracy $\delta=1/\sqrt{K_p^p F(\mu)}$. Inclusion of $\pi(\mu)$ and $\sqrt{F(\mu)}$ help reduce the measure of uncertainty, their ratio is dimension free and invariant to reparameterization (Wallace and Dowe, 1993). Since MML is a Bayesian method and depends on the choice of prior density of the parameters, there is scope in selecting the prior. As a result, estimators and tests based on the message length may be different for different priors.

As mentioned by Wallace and Freeman (1987), the MML principle was possibly first initiated by Solomonoff (1964) as a general principal of inductive inference. This principle was applied in a series of papers by Boulton (1975), Boulton and Wallace (1969, 1970, 1973, 1975) and Wallace and Boulton (1968). Their main concern was the application of the MML principle in estimation and model selection for intrinsic classification problems (Wallace, 1986, 1990 and Wallace and Boulton, 1968) as a computer based method. Recently, Wallace and Freeman (1987) advanced the idea of MML as an alternative method of estimation and test construction. Also, 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 ML estimator. Following this, Wallace and Dowe (1993) applied the MML approach to estimating the von Mises concentration parameter and observed its improved accuracy over the ML estimator for small sample sizes. Also, Wallace and Dowe (1994) provided a brief overview of message length based estimation and the application of the message length intrinsic classification programme, SNOB.

For model (1), an approximate message length given by Wallace and Freeman (1987) and accurate to $\delta = 1/\sqrt{K_s^s F(\theta,\sigma^2,\beta)}$ is

$$-\log\left[\frac{\pi(\theta,\sigma^2,\beta)L(\theta,\sigma^2,\beta)}{\sqrt{F(\theta,\sigma^2\beta)}}\right]+\frac{s}{2}(1+\log K_s) \tag{24}$$

where $\pi(\theta,\sigma^2,\beta)$ is a prior density for $\gamma = (\theta',\sigma^2,\beta',)'$, $F(\theta,\sigma^2,\beta)$ is the determinant of the information matrix and $s = k+p+1$. In this section, we assume that θ is a scalar so $p = 1$. Using the results of section 2.4.1,

$$\frac{1}{2}\log F(\theta,\sigma^2,\beta) = -\frac{k+2}{2}\log\sigma^2+\frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right|$$

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

Assuming the non-informative prior $\pi(\theta,\sigma^2,\beta)=\frac{1}{\sigma^2}$, the message length function (24) becomes

$$ML = \frac{m}{2}\log\sigma^2+\frac{1}{2}\log|\Omega(\theta)|+\frac{1}{2\sigma^2}(y-X\beta)'\Omega(\theta)^{-1}(y-X\beta)+\frac{1}{2}\log\left|X'\Omega(\theta)^{-1}X\right|$$

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

$$+\frac{s}{2}(1+\log K_s)-\log 2. \tag{25}$$

3 Estimation of Error Variance Using Different Methods

In this section, we use each method to find the estimator of the error variance for the classical linear regression model. This may help us understand the relative strengths of the different likelihood based approaches, given this is a situation in which the classical ML estimator is known to be biased.

Consider the special case of (1),

$$y = X\beta+u,\ u\sim N(0,\sigma^2 I). \tag{26}$$

The log likelihood of (26) ignoring constant terms is

$$l(y;\sigma^2,\beta) = -\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta) \tag{27}$$

and the profile (or concentrated) likelihood for σ^2 is

$$l_p(y;\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}y'My \tag{28}$$

where $M = I_n - X(X'X)^{-1}X'$. Differentiating $l_p(y;\sigma^2)$ with respect to σ^2 and equating to zero we get

$$\hat{\sigma}^2 = \frac{y'My}{n},$$

which is the familiar biased estimator of the error variance. In the subsequent subsections, the estimator of the error variance by the different approaches is provided.

3.1 Marginal Likelihood

We need to derive the marginal likelihood of σ^2 from model (26). Let $\hat{\beta}$ and $z = My$ be the OLS estimator of β and residual vector from (26) respectively and let L be the $n\times m$ matrix defined for equation (6) so that $L'M = L'$. Then, consider the one-to-one transformation of y to $\hat{\beta}$ and $L'z$ and observe that

$$\hat{\beta} \sim N(\beta,\sigma^2(X'X)^{-1}) \text{ and } L'z \sim N(0,\sigma^2 I_m).$$

Note that $L'z$ is $m\times 1$ and independent of $\hat{\beta}$ so through this transformation, the likelihood for model (26) can be written as the product of the density for $\hat{\beta}$ and the density for $L'z$. The latter is the marginal likelihood for σ^2 because it does not contain β and there is no loss of information about σ^2. In log form and ignoring constant terms, the marginal likelihood is

$$l_m(y;\sigma^2) = -\frac{m}{2}\log\sigma^2 - \frac{y'My}{2\sigma^2}. \tag{29}$$

Differentiating $l_m(y;\sigma^2)$ with respect to σ^2 and equating to zero we get

$$\hat{\sigma}^2 = \frac{y'My}{m},$$

which is the familiar unbiased estimator of the error variance.

3.2 Modified Profile Likelihood

The log of the MPL for model (26) is given by

$$l_{mp}(y;\sigma^2) = l_p(y;\sigma^2) + \log\left|\frac{\partial\hat{\beta}}{\partial\hat{\beta}_{\sigma^2}}\right| - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\sigma^2,\beta)}{\partial\beta\partial\beta'}\right|_{\hat{\beta}} \tag{30}$$

where $\hat{\beta}$ is the ML estimator of β and $\hat{\beta}_{\sigma^2}$ is the ML estimator of β for fixed σ^2 from (26). The parameters β and σ^2 are orthogonal, so $\left|\frac{\partial\hat{\beta}}{\partial\hat{\beta}_{\sigma^2}}\right| = 1$ and (30) reduces to

$$l_{mp}(y;\sigma^2) = l_p(y;\sigma^2) - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\sigma^2,\beta)}{\partial\beta\partial\beta'}\right|_{\hat{\beta}}.$$

From (26), $-\left.\frac{\partial^2 l(y;\sigma^2,\beta)}{\partial\beta\partial\beta'}\right|_{\hat{\beta}} = \frac{X'X}{\sigma^2}$ so that

$$l_{mp}(y;\sigma^2) = -\frac{m}{2}\log\sigma^2 - \frac{1}{2}\log|X'X| - \frac{y'My}{2\sigma^2}.$$

Thus, for this model, the marginal likelihood and the MPL are equivalent and therefore give the same unbiased estimator for σ^2.

3.3 Conditional Profile Likelihood

The information matrix for model (26) is

$$\begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{X'X}{\sigma^2} \end{bmatrix}$$

which indicates that σ^2 and β are orthogonal and satisfy condition (37) of section 4 below. This implies that the MPL and the CPL are equivalent in the moderate-derivative sense (Barndorff-Nielsen and McCullagh, 1993; see section 4.1 below). This fact implies that the estimator of σ^2 using this approach is the same as that of the MPL approach.

3.4 Expected Log-Likelihood Approach

In this approach, an estimator of σ^2 can be obtained by equating

$$\frac{\partial}{\partial\sigma^2} l_p(y;\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} y'My \tag{31}$$

to its expectation. Taking the expectation of (31), we get $-\frac{k}{2\sigma^2}$. The estimator of σ^2 is the solution for σ^2 from the estimating equation

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} y'My = -\frac{k}{2\sigma^2},$$

which again gives $\hat{\sigma}^2 = \frac{y'My}{m}$.

3.5 Adjusted Profile Likelihood

The score function $S(\sigma^2)$ is given by

$$S(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} y'My$$

and the adjusted score function is

$$\tilde{S}(\sigma^2) = W(\sigma^2)\{S(\sigma^2) - m(\sigma^2)\}$$

where $m(\sigma^2) = E_{\sigma^2,\hat{\beta}_{\sigma^2}}(S(\theta))$

and

$$W(\sigma^2) = [\mathrm{var}_{\sigma^2,\hat{\beta}_{\sigma^2}}\{S(\sigma^2)\}]^{-1}\left[-E_{\sigma^2,\hat{\beta}_{\sigma^2}}\left(\frac{\partial S(\sigma^2)}{\partial\sigma^2}\right) + \frac{\partial m(\sigma^2)}{\partial\sigma^2}\right].$$

Now

$$m(\sigma^2) = -\frac{k}{2\sigma^2}, \quad \frac{\partial m(\sigma^2)}{\partial\sigma^2} = \frac{k}{2\sigma^4}, \quad \frac{\partial S(\sigma^2)}{\partial\sigma^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} y'My,$$

$$-E_{\sigma^2,\hat{\beta}}\left(\frac{\partial S(\sigma^2)}{\partial\sigma^2}\right) = -\frac{n}{2\sigma^4} + \frac{m}{\sigma^4}$$

and

$$\mathrm{var}_{\sigma^2,\hat{\beta}_{\sigma^2}}\{S(\sigma^2)\} = \frac{m}{2\sigma^4}.$$

Thus $W(\sigma^2) = \left\{-\frac{n}{2\sigma^4} + \frac{m}{\sigma^4} + \frac{k}{2\sigma^4}\right\}\frac{2\sigma^4}{m} = 1.$

The adjusted score function is

$$\tilde{S}(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{y'My}{2\sigma^4} + \frac{k}{2\sigma^2}.$$

The estimator of σ^2 can be obtained by equating the above adjusted score function to zero, which gives

$$\hat{\sigma}^2 = \frac{y'My}{m},$$

the familiar unbiased estimator.

3.6 Average Likelihood

To find the posterior density for σ^2, we need to integrate the r^{th} power of $L(y;\sigma^2,\beta)$ with respect to β. Using the non-informative prior $\pi(\beta,\sigma) = \frac{1}{\sigma}$ for β and σ, the average likelihood for σ^2 is given by

$$L^A(y;\sigma^2) = 2^{-\frac{k}{2}} \frac{1}{\sigma^n} \exp[-\frac{y'My}{2\sigma^2}] \tag{32}$$

which is directly proportional to the profile likelihood. In this case, the estimator of the error variance is not unbiased.

3.7 Approximate Conditional Profile Likelihood

For model (26), likelihood function (21) becomes

$$l_{cpo}(y;\sigma^2) = l_p(y;\sigma^2) - \frac{1}{2}\log\left|-\frac{\partial^2 l(y;\sigma^2,\beta)}{\partial\beta\partial\beta'}\right|_{\hat{\beta}}.$$

The last term in (21) disappears in this case because σ^2 and β are orthogonal. Finally ignoring constants, the log likelihood function is given by

$$l_{cpo}(y;\sigma^2) = -\frac{m}{2}\log\sigma^2 - \frac{1}{2}\log|X'X| - \frac{y'My}{2\sigma^2}. \tag{33}$$

The estimator of σ^2 from (33) is the familiar unbiased estimator.

3.8 Minimum Message Length

For model (26), (24) can be written as

$$-\log\left[\frac{\pi(\sigma^2,\beta)L(y;\sigma^2,\beta)}{\sqrt{F(\sigma^2,\beta)}}\right] + \frac{D}{2}(1+\log K_D) \tag{34}$$

where $\pi(\sigma^2,\beta)$ is the prior density for $\gamma_1 = (\sigma^2,\beta')'$, $L(y;\sigma^2,\beta)$ is the likelihood function for model (26), $F(\sigma^2,\beta)$ is the determinant of the information matrix, and K_D is the $D = k + 1$ dimensional optimal quantizing lattice constant as defined in section 2.11. Using the results of section 3.3

$$\frac{1}{2}\log F(\sigma^2,\beta) = -\frac{k+2}{2}\log\sigma^2 + \frac{1}{2}\log|X'X| + \frac{1}{2}\log\frac{n}{2}.$$

Using the non-informative prior $\pi(\sigma^2,\beta) = \frac{1}{\sigma^2}$ for the parameters (σ^2,β'), (26) becomes

$$ML = \frac{m}{2}\log\sigma^2 + \frac{1}{2}\log|X'X| + \frac{1}{2\sigma^2}u'u + \frac{1}{2}\log\frac{n}{2} + \frac{D}{2}(1+K_D).$$

The MML estimators of the parameter β and σ^2 are

$$\hat{\beta} = (X'X)^{-1}X'y \text{ and } \hat{\sigma}^2 = \frac{y'My}{m},$$

respectively. Thus the MML estimator of σ^2 is unbiased for our choice of prior but for any other choice of prior this estimator will be biased.

4 Comparison of Likelihood and Related Methods

All the different likelihood functions are designed to deal with nuisance parameters, although the manner in which they do this is different. The approaches suggested by Kalbfliesch and Sprott (1970) and Cox and Reid (1987) appear to be the most popular. Many researchers have used these two likelihood functions for estimation and testing problems. The other approaches have limited applicability in econometric analysis. We will now discuss how the different likelihood functions differ.

4.1 Theoretical Comparisons

Barndorff-Nielsen and McCullagh (1993) investigated the relationship between the profile likelihood, CPL and the MPL. For model (1), the relationship is

$$L_{mp}(y;\theta) = D_1(\theta)L_{cp}(y;\theta) = D_1(\theta)\left\{\left|-\frac{\partial^2 l(y;\theta,\sigma^2,\beta)}{\partial\gamma^{\dagger}\partial\gamma^{\dagger\prime}}\right|_{\hat{\gamma}_\theta^{\dagger}}\right\}^{-\frac{1}{2}} L_p(y;\theta) \qquad (35)$$

where $\gamma^{\dagger}$ and $\hat{\gamma}_{\theta}^{\dagger}$ are defined in section 2.3 and $D_1(\theta) = \left|\partial\hat{\gamma}^{\dagger} / \partial\hat{\gamma}_{\theta}{}^{\dagger}\right|$. In order to identify situations in which the CPL is equivalent to the MPL, the following conditions were examined by Barndorff-Nielsen and McCullagh (1993):

$$D_1(\theta) = 1 \tag{36}$$

and more generally

$$D_1(\theta) = 1 + O(n^{-1}). \tag{37}$$

They explained that the CPL and the MPL are equivalent in the large-deviation sense if (37) holds for $\theta - \hat{\theta} \doteq O(1)$ and they are equivalent in the weaker moderate-derivative sense if (37) holds for $\theta - \hat{\theta} = O(n^{-\frac{1}{2}})$ where $\hat{\theta}$ is the ML estimator of θ. For model (26), the marginal log likelihood for σ^2 is given by (29). Barndorff-Nielsen (1988) showed (after a Laplace approximation) that:

$$L_m(y;\sigma^2) \approx L_p(y;\sigma^2)\left\{\left|-\frac{\partial l(y;\sigma^2,\beta)}{\partial\beta\partial\beta'}\right|_{\hat{\beta}}\right\}^{-\frac{1}{2}} \tag{38}$$

which is an approximation to the CPL in the (σ^2,β) parameterization and $L_m(y;\sigma^2)$ is the marginal likelihood for σ^2. In this case $D_1(\sigma^2) = 1$ (normal linear regression), so that the CPL and the MPL are approximations to the marginal likelihood.

Many authors have tried to compare different likelihood functions. In this context, Cruddas, Reid and Cox (1989) observed on the basis of a simulation study that CPL and marginal likelihood are the same for the standardized residuals of short Gaussian first-order autoregressive processes with different means but common correlation and variance. Moreover, Bellhouse (1990) found that the CPL and marginal likelihood are equivalent for correlated parameters in a general normal regression model. Furthermore, Reid (1995) noted that the CPL is not invariant under one-to-one reparameterizations of the nuisance parameter σ^2 that leave the parameter of interest fixed. She pointed out that this lack of invariance can be avoided by using the MPL.

The marginal likelihood is given by (5) and $\hat{s}(X'\Omega(\theta)^{-1}X)^{-1}$ is proportional to the estimated variance-covariance matrix of the ML estimator of β for given $\Omega(\theta)$. Then (5) can be written as

$$L_m(y;\theta) = \frac{[\text{est var}(\hat{\beta})]^{1/2}}{\hat{s}^{n/2}|\Omega(\theta)|^{1/2}}. \tag{39}$$

Bellhouse (1990) observed that using the transformation $\lambda = \log\sigma + \frac{1}{2}\log|\Omega(\theta)|/n$ in (2) and keeping β the same, the CPL for θ is also given by (5), provided that λ and θ, and β and θ are orthogonal. Tunnicliffe Wilson (1989) noted that if σ^2 is parameterized as e^{ψ} and ψ includes in the parameter set β of the CPL, then it gives the marginal likelihood. Laskar and King (1998) showed that the marginal likelihood and CPRL are equivalent via the relationship

$$\bar{l}^{*}_{cpr}(y;\theta) = \frac{m-2}{m} l_m(y;\theta)$$

so, for the purpose of estimating θ, the marginal likelihood and CPRL are equivalent but this is not necessarily true for likelihood based tests of θ.

There are some similarities between the message length function in (25) and the marginal likelihood in (5). Returning to the case of $p = 1$, the MML estimators of β and σ^2 conditional on θ are

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

Putting these estimators in (25), we get, ignoring constant terms,

$$ML(y;\theta) = \frac{m}{2}\log\hat{\sigma}^2 + \frac{1}{2}\log|\Omega(\theta)| + \frac{1}{2}\log|X'\Omega(\theta)^{-1}X|$$

$$+\frac{1}{2}\log\left(n \times tr\left[-\frac{\partial\Omega(\theta)}{\partial\theta}\frac{\partial\Omega(\theta)^{-1}}{\partial\theta}\right] - \left\{tr\left[\Omega(\theta)^{-1}\frac{\partial\Omega(\theta)}{\partial\theta}\right]\right\}^2\right). \tag{40}$$

The form (40) is close to the marginal likelihood for θ; the only difference is the last term in (40).

It can be concluded from the above discussion that in some situations $l_m(y;\theta)$, $l_{mp}(y;\theta)$ and $l_{cp}(y;\theta)$ are equivalent. Ara and King (1993) demonstrated that $l_m(y;\theta)$, and the log likelihood of the maximal invariant statistic are the same. To our knowledge, there is no available literature on computing adjusted profile likelihoods, canonical likelihoods, expected likelihoods, and ACPL. This is clearly a shortcoming of the literature to date.

Mahmood and King (1997) investigated the unbiasedness property of score vectors for different likelihood functions considered in section 2. They observed that score vectors based on the marginal likelihood and CPRL are unbiased. This means that expected values of the score functions based on these two likelihoods are zero. Possibly, Godambe (1960) first introduced the concept of unbiased estimating equations to demonstrate an optimum property of regular ML estimation and subsequently it was applied and extended by Godambe and Thompson (1974), Ferreira (1982), Chandrasekar and Kale (1984) and Conniffe (1990a) for different estimation problems. It can be shown that the score vector for the message length function derived in section 2.11 is biased. The LM test based on unbiased score vectors can have good small sample properties (Ara and King 1993, Ara 1995 and Laskar and King 1998) and those based on biased score vectors can have poor small sample properties (King, 1987, Honda, 1988 Moulton and Randolph, 1989 and Laskar and King, 1997b).

4.2 Empirical Comparisons

Many researchers have empirically investigated the performance of different likelihood and related methods in terms of estimation and testing of θ in different contexts. In this section, we review this literature.

4.2.1 Estimation

Cooper and Thompson (1977) applied marginal likelihood estimation to time series models and investigated the small sample properties of the estimator for the parameter of the MA(1) model. They reported a significant reduction in bias of the estimator compared to that of the classical likelihood function. Also Grose (1992) used the marginal likelihood for estimating the coefficient of lagged dependent variable in the dynamic regression model. She reported that the estimator based on marginal likelihood is less biased compared to the OLS estimator. Wallace and Freeman (1992) applied the MML approach to the problem of estimating the parameters of a multivariate Gaussian model in which they modelled the correlation structure by a single common factor. They found that MML estimates on average are more accurate than those of the ML estimator in terms of estimating both the factor

loadings and the factor scores, if the former exist. Furthermore, Wallace and Dowe (1993) applied the MML approach to estimating the von Mises concentration parameter and observed its improved accuracy over the ML estimator for small sample sizes.

Laskar and King (1996) investigated the small sample properties of six different MML estimators in the context of (1) and MA(1) and AR(1) regression disturbances. They summarized the results using the loss function, |bias| + $\frac{1}{\lambda}$(standard deviation) + $\frac{1}{\lambda^2}$|skewness| + $\frac{1}{\lambda^3}$|kurtosis - 3| where $\lambda = 3$. The loss function is dominated by the bias and standard deviation terms. In a lot of cases, the estimators are nearly unbiased so the dominant term was the standard deviation. They reported that the estimators based on the combination of parameter orthogonality and message length function are closer to normal for moderate and small values of $|\theta|$. However, for other values of θ away from 0, the performance of ordinary message length functions based estimators are relatively better. Their findings showed that for estimating the MA(1) and AR(1) disturbances parameter, message length functions obtained by combining MML with CPL 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. Recently, Dowe and Wallace (1997) resolved the Neyman-Scott problem by using MML principle. They considered multiple Gaussian distributions with unknown means and identical but unknown standard deviation and observed that the ML estimator of variance is inconsistent but the MML estimator of variance is consistent.

Laskar and King (1998) also investigated the small sample properties of the ML estimator of θ in the context of (1) and MA(1) and AR(1) regression disturbances based on the (i) profile likelihood, (ii) marginal likelihood, (iii) CPL, (iv) CPRL and (v) ACPL. They concluded that the distribution of the estimators based on the marginal likelihood, CPRL and CPL are closer to the normal distribution on the basis of their respective bias, standard deviation, skewness and kurtosis. Laskar and King's results reflect that CPL based estimators typically have the smallest average loss compared to those based on the profile likelihood and other modified likelihood functions. Also Laskar (1998) constructed confidence intervals using different

modified likelihood functions in order to compare their small sample properties in the context of AR(1) linear regression disturbances. They reported that the marginal likelihood based confidence intervals are best and the CPRL and ACPL based confidence intervals are the second best.

4.2.2 Testing

Cordus (1986) showed, through a simulation study for the classical linear regression model with AR(1) disturbances, that the use of the marginal likelihood improves the LR test. She derived a test statistic based on the OLS residuals which is a modification of the score statistic and observed that the resulting test performs better than the classical LR test. Also, Cruddas et al. (1989) undertook a simulation study to find confidence intervals for the correlation parameter, based on observations from a large number of short Gaussian AR(1) processes with different means but common correlation and variance. They reported better small sample properties of the CPL based estimators and LR tests than those of the classical likelihood.

Mukerjee (1993) constructed the LR test for a general multiparameter set-up based on the CPL and indicated its superiority over the usual LR test. He also showed that the use of the adjusted profile likelihood can improve the LR test and exemplified this in the cases of (a) parameter orthogonality and (b) no parameter orthogonality.

Ara and King (1993) derived general formulae for the LR, LM, Wald and ALMMP tests for linear regression disturbances using the marginal likelihood and investigated the small sample properties of these tests for testing the parameters of fourth-order autoregressive disturbances and the presence of Hildreth-Houck random coefficients. They reported better small sample sizes of these tests compared to those based on the classical likelihood. Their study also reported better centred power curves of all the marginal likelihood based tests. In addition, Ara and King (1995) investigated the small sample sizes and powers of the LR, LM, Wald and ALMMP tests for a subvector of the parameter vector θ based on the marginal likelihood. They reported better improvements in small sample sizes and powers of the marginal likelihood based tests compared to those of Ara and King (1993). This significant improvement of small sample sizes and powers of the tests occurred due to better handling of nuisance parameters.

Laskar and King (1997b) investigated the small sample properties of the LR, LM, Wald and Null Wald (NW, Laskar and King (1997a)) tests in the context of (1) and MA(1) and AR(1) regression disturbances based on the classical likelihood and six different message length functions. They reported that all the tests based on simple message length functions for MA(1) processes and those based on combinations of parameter orthogonality and message length functions for AR(1) processes have better small sample sizes which are closer to their asymptotic size and their power curves are better centred. They also reported that in general, sizes of all the message length based LM tests are significantly higher than the asymptotic size. Because of the unbiasedness of score functions (Mahmood and King, 1997) of all message length functions, sizes of their LM tests are very poor, and away from the asymptotic size. Also, Grose (1997, 1998) constructed the standard t-test, LR, LM and Wald tests based on marginal likelihood and profile likelihood for the coefficient of lag dependent variable in the first order dynamic regression model and investigated their small sample properties. She reported that for positive values of the coefficient, sizes and powers of all the tests based on marginal likelihood have better small sample properties compared to those based on profile likelihood but for negative values of the coefficient, the marginal likelihood based tests do not perform so well.

Rahman and King (1998) developed the marginal likelihood based LM and ALMMP tests for situations in which the parameter vector of the error structure is partitioned into two parts, with one being the parameters of interest and the other being nuisance parameters. For this testing problem, all nuisance parameters cannot be eliminated using the likelihood of the maximal invariant or the marginal likelihood. Instead, they constructed tests, in which those nuisance parameters which could not be eliminated were replaced by their maximum marginal likelihood estimators. They observed a higher level of improvement in both sizes and powers, particularly for the LM test, than that reported by Ara and King (1993). It has been found that the maximum marginal likelihood based LM test can improve both the small sample size and power relative to that of the conventional LM test

Laskar and King (1998) investigated the small sample properties of the LR, LM and Wald tests in the context of (1) and MA(1) and AR(1) regression disturbances based on different modified likelihood functions mentioned in section 4.1. They

reported that the sizes of all the tests based on marginal likelihood, CPL and CPRL are closer to the nominal size compared to their classical counterparts. The powers of all the tests based on modified likelihoods are better centred and less biased than those based on the classical likelihood. Sizes of the marginal likelihood based LM tests are most impressive with almost all of them being closest to the asymptotic size. In this regard, Mahmood and King (1997) observed that the score function based on marginal likelihood and CPRL are unbiased and the LM test based on an unbiased score function can have best small sample properties.

Laskar (1998) investigated and compared the small sample properties of estimators and tests based on eleven different likelihood and message length functions in the context of MA(1) and AR(1) regression disturbances. He found that overall the marginal likelihood is best for testing while the CPL is best for estimation. He also mentioned that for MA(1) disturbances, sizes of the Wald, alternative Wald and null Wald tests are more accurate when modified likelihood functions are replaced by message length functions. These results may be caused by the identification problem for MA(1) disturbances. If there is a problem of lack of identification, the information matrix reacts through the $F(\mu)$ term in (23) and helps solve this problem (Martin, 1997). It seems that the inclusion of this factor may help overcome the side-effects of the identification problem. Consequently all versions of message length based Wald tests have better small sample sizes compared to those based on modified likelihood functions.

4.2.3 Model Selection

Tunnicliffe Wilson (1989) argued that the application of the marginal likelihood in time series regression has significant effects on model selection. He used the smallest residual variance as a selection criteria for different models using profile likelihood and marginal likelihood. Also, Grose and King (1993) proposed the use of the marginal likelihood for the problem of selecting between AR(1) or MA(1) regression disturbances in (1). Via a simulation study, they found that the presence of nuisance parameters can seriously affect the probabilities of correct selection. They used Monte Carlo methods to find more appropriate penalties and found that the application of information criteria to marginal likelihoods rather than classical

likelihoods gives improved small sample selection probabilities. Recently, Baxter and Dowe (1996) applied the MML criteria for choosing the degree of a polynomial in least-squares regression and reported that it selects the degree of polynomial accurately compared to minimum description length (MDL) (Rissanen, 1978), Akaike's information criterion (AIC) and consistent AIC with Fisher information (CAICF) (Bozdogan, 1987). More recently, Oliver and Forbes (1997) developed the Bayes Factor and MML approach of segmenting a time series and compared them with AIC, MDL and Bayesian information criteria (BIC) using Monte Carlo simulations. They report that the MML performs better than all other criteria.

4.2.4 Forecasting

Latif and King (1993) introduced a new approach for time-series forecasting based on the linear regression model in the presence of AR(1) disturbances. They suggested a weighted average of predictions, assuming different values of the AR(1) parameter with weights proportional to the marginal likelihood of that parameter. Their simulation results show that the new approach can produce better forecasts compared to existing procedures, which is a consequence of the application of the marginal likelihood.

5 Concluding Remarks

Inference for a parameter of interest in the presence of nuisance parameters is a long-standing problem for statisticians and econometricians. As a result, many authors have attempted to modify the likelihood function in several ways in order to provide a satisfactory way of handling this problem. This paper has discussed eleven likelihood and related methods which are available for such purposes. For the simple linear regression model, the ML estimator of the error variance is unbiased in all the approaches except for the profile likelihood and average likelihood.

Marginal likelihood is a popular approach for making inference about the parameter of interest. Unfortunately, the marginal likelihood cannot be constructed in all situations and REML applies only to the disturbance parameters in the linear model. As an alternative, Barndorff-Nielsen (1983) proposed the MPL and Cox and Reid (1987) introduced the idea of CPL. CPL needs the parameter(s) of interest and

the nuisance parameters to be orthogonal. If orthogonality does not exist, Cox and Reid suggested reparameterizing the nuisance parameters to get the required orthogonality. The major drawback of CPL is the nonuniqueness of the orthogonal parameterization and the fact that it may not always be possible to find an orthogonal parameterization. These two problems were partially resolved by Cox and Reid (1993). They derived an approximation to the CPL which depends on the ML estimator from the profile likelihood. Cox and Reid (1987) mentioned the difficulty of using the MPL because it requires conditioning on an appropriate ancillary statistic.

As an alternative to modified likelihoods, message length is a Bayesian method, which contains all the parameters of model (1) and has the usual consequence of nuisance parameters. There is also evidence which suggests that, where possible, information criteria model selection procedures such as AIC or BIC should be based on modified likelihoods rather than classical likelihoods.

In conclusion, the discussion presented in this paper, indicates that the CPL is the best for estimation and marginal likelihood is the best for testing the parameter of interest after eliminating the effect of nuisance parameters while if there is a minor identification problems, message length functions can play a positive role in Wald tests at least for the general linear regression model.

Acknowledgement

This research has been supported by an ARC grant. We are grateful for comments from and discussions with David L. Dowe, Gale M. Martin, Ismat Ara, Shafiqur Rahman, David Harris and Vincenzo Matassa.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society* B 53, 111-142.

Aitkin, M. (1993). Posterior Bayes factor analysis for an exponential regression model. *Statistics and Computing* 3, 17-22.

Ara, I. (1995). Marginal likelihood based tests of regression disturbances. Unpublished Ph.D. thesis, Monash University.

Ara, I. and M.L. King (1993). Marginal likelihood based tests of regression disturbances. Mimeo (Monash University).

Ara, I. and M.L. King (1995). Marginal likelihood based tests of a subvector of the parameter vector of linear regression disturbances. In C.S. Forbes, P. Kofman and T.R.L. Fry (eds.), *Proceedings of the Econometrics Conference at Monash University*, 69-106.

Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343-365.

Barndorff-Nielsen, O.E. (1988). *Parametric Statistical Models and Likelihoods*. Lecture Notes in Statistics 50. Heidelberg: Springer-Verlag.

Barndorff-Nielsen, O.E. and P. McCullagh (1993). A note on the relation between modified profile likelihood and the Cox-Reid adjusted profile likelihood. *Biometrika* 80, 321-328.

Baxter, R.A. and D.L. Dowe (1994). Model selection in linear regression using the MML criterion. In J.A. Storer and M. Cohn (eds.), *Proceedings of the 4th IEEE Data Compression Conference*. IEEE Computer Society Press, Los Alamitos, CA, America, 498.

Bellhouse, D.R. (1978). Marginal likelihood methods for distributed lag models. *Statistiche Hefte* 19, 2-14.

Bellhouse, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika* 77, 743-746.

Bewley, R. (1986). *Allocation Models: Specification, Estimation and Applications*. Ballinger, Boston.

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

Boulton, D.M. and C.S. Wallace (1969). The information content of a multistate distribution. *Journal of Theoretical Biology* 23, 269-278.

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.

Boulton, D.M. and C.S. Wallace (1975). An information measure for single-link classification. *The Computer Journal* 18, 236-238.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52, 345-370.

Chandrasekar, B. and B.K. Kale (1984). Unbiased statistical estimation functions for parameters in presence of nuisance parameters. *Journal of Statistical Planning and Inference* 9, 45-54.

Chesher, A. and G. Austin (1991). The finite sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics* 47, 153-173.

Cochrane, D. and G.H. Orcutt (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association* 44, 32-61.

Conniffe, D. (1987). Expected maximum likelihood estimation. *The Statistician* 36, 317-329.

Conniffe, D. (1988). Obtaining expected maximum log likelihood estimators. *The Statistician* 37, 441-449.

Conniffe, D. (1990a). Testing hypotheses with estimated scores. *Biometrika* 77, 97-106.

Conniffe, D. (1990b). Applying estimated score tests in econometrics. Paper presented at the 1990 World Congress of the Econometric Society, Barcelona.

Conway, J.H. and N.J.A. Sloan (1988). *Sphere Packings, Lattices and Groups.* Springer-Verlag, London.

Cooper, D.M. and R. Thompson (1977). A note on the estimation of parameters of the autoregressive-moving average process. *Biometrika* 64, 625-628.

Cordus, M. (1986). The use of the marginal likelihood in testing for serial correlation in time series regression. Unpublished M. Phil. thesis, University of Lancaster.

Cox, D.R. (1988). Some aspects of conditional and asymptotic inference: A review. *Sankhya* 50, 314-337.

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.

Cox, D.R. and N. Reid (1993). A note on the calculation of adjusted profile likelihood. *Journal of the Royal Statistical Society* B 55, 467-471.

Cruddas, A.M., N. Reid and D.R. Cox (1989). A time series illustration of approximate conditional likelihood. *Biometrika* 76, 231-237.

Dowe, D.L. and C.S. Wallace (1997). Resolving the Neyman-Scott problem by minimum message length. Technical Report No.97/307, Monash University.

Ferguson, H. (1992). Asymptotic properties of a conditional maximum-likelihood estimator. *The Canadian Journal of Statistics* 20, 63-75.

Ferguson, H., N. Reid and D.R. Cox (1991), Estimating equations from modified profile likelihood. In V.P. Godambe (eds.), *Estimating Functions*. Oxford University Press, 279-293.

Ferreira, P.E. (1982). Multiparametric estimating equations. *Annals of the Institute of Statistical Mathematics* 34A, 423-431.

Fraser, D.A.S. (1967). Data transformations and the linear model. *Annals of Mathematical Statistics* 38, 1456-1465.

Fraser, D.A.S. and N. Reid (1989). Adjustments to profile likelihood. *Biometrika* 76, 477-488.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* 31, 1208-1212.

Godambe, V.P. and M.E. Thompson (1974). Estimating equations in the presence of a nuisance parameter. *The Annals of Statistics* 2, 568-571.

Grose, S.D. (1992). Marginal likelihood estimation in the dynamic linear regression model. Paper presented at the 1992 Australasian Meeting of the Econometric Society, Monash University, Melbourne.

Grose, S.D. (1997). Marginal likelihood based testing in the dynamic linear model. In P. Bardsley and V.L. Martin (eds.), *Proceedings of the Econometric Society Australasian Meeting* 2. University of Melbourne, 177.

Grose, S.D. (1998). Marginal likelihood methods in econometrics. Unpublished Ph.D. thesis Monash University.

Grose, S.D. and M.L. King (1993). The use of information criteria for model selection between models with equal numbers of parameters. Mimeo (Monash University).

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* 61, 383-385.

Hinde, J. and M. Aitkin (1987). Canonical likelihoods: A new likelihood treatment of nuisance parameters. *Biometrika* 74, 45-58.

Honda, Y. (1988). A size correction to the Lagrange multiplier test for heteroscedasticity. *Journal of Econometrics* 38, 375-386.

Huzurbazar, V.S. (1950). Probability distributions and orthogonal parameters. *Proceedings of Cambridge Philosophical Society* 46, 281-284.

Kalbfleisch, J.D. and D.A. Sprott (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society* B 32, 175-208.

King, M.L. (1987). Testing for autocorrelation in linear regression models: A survey. In M.L. King and D.E.A. Giles (eds.), *Specification Analysis in the Linear Model.* Routledge and Kegan Paul, London, 19-73.

King, M.L. (1996). Hypothesis testing in the presence of nuisance parameters. *Journal of Statistical Planning and Inference* 50, 103-120.

King, M.L. and M. McAleer (1987). Further results on testing AR(1) against MA(1) disturbances in the linear regression model. *Review of Economic Studies* 54, 649-663.

King, M.L. and P.X. Wu (1997). Locally optimal one-sided tests for multiparameter hypotheses. *Econometric Reviews* 16, 131-156.

Laskar M.R. (1998). Estimation and testing of linear regression disturbances based on modified likelihood and message length functions. Unpublished Ph.D. thesis, Monash University.

Laskar, M.R. and M.L. King (1996). Estimation of regression disturbances based on minimum message length. In D.L. Dowe, K.B. Korb and J.J. Oliver (eds.), *Proceedings of the Conference ISIS'96: Information, Statistics and Induction in Science,* World Scientific, Singapore, 92-101.

Laskar, M.R. and M.L. King (1997a), Modified Wald test for regression disturbances, *Economics Letters* 56, 5-11.

Laskar, M.R. and M.L. King (1997b). Testing of regression disturbances based on minimum message length. In P. Bardsley and V.L. Martin (eds.), *Proceedings of the Econometric Society Australasian Meeting* 2, University of Melbourne, 179-202.

Laskar, M.R. and M.L. King (1998). Estimation and testing of regression disturbances based on modified likelihood functions. *Journal of Statistical Planning and Inference*, forthcoming.

Latif, A. and M.L. King (1993). Linear regression forecasting in the presence of AR(1) disturbances. *Journal of Forecasting* 12, 513-524.

Levenbach, H. (1972). Estimation of autoregressive parameters from a marginal likelihood function. *Biometrika* 59, 61-71.

Macaskill, G.T. (1993). A note on adjusted profile likelihoods in non-linear regression. *Journal of the Royal Statistical Society* B 55, 125-131.

Mahmood, M. and M.L. King (1997). Modified likelihood functions and the biasedness of their estimating equations. Mimeo (Monash University).

Martin, G.M. (1997). Fractional cointegration: Bayesian inference using Jeffreys prior. Working paper No. 1/98, Department of Econometrics and Business Statistics, Monash University.

McCullagh, P. and R. Tibshirani (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society* B 52, 325-344.

Moulton, B.R. and W.C. Randolph, (1989). Alternative tests of the error components model. *Econometrica* 57, 685-693.

Mukerjee, R. (1993). An extension of the conditional likelihood ratio test to the general multiparameter case. *Annals of Institute of Statistical Mathematics* 45, 759-771.

Neyman, J. and E.L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1-32.

Oliver, J.J. and C.S. Forbes (1997). Bayesian approaches to segmenting a simple time series. Working paper No. 14/97, Department of Econometrics and Business Statistics, Monash University.

Patterson, H.D. and R. Thompson (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* 58, 545-554.

Rahman, S. and M.L. King (1998). Marginal likelihood score based tests of regression disturbances in the presence of nuisance parameters. *Journal of Econometrics* 82, 81-106.

Reid, N. (1995). The role of conditioning in inference. *Statistical Science* 10, 138-199.

Simonoff, J.S. and C. Tasi (1994). Use of modified profile likelihood for improved tests of constancy of variance in regression. *Applied Statistics* 43, 357-370.

Solomonoff, R. (1964). A formal theory of inductive inference I, II. *Information and Control* 7, 1-22, 224-254.

Thompson, R. (1973). The estimation of variance and covariance components with an application when records are subjected to culling. *Biometrics* 29, 527-550.

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.

Verbyla, A.P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* 32, 227-230.

Wallace, C.S. (1986). An improved program for classification. *9th Australian Computer Science Conference* (*ACSC-9*) 8, 357-366.

Wallace, C.S. (1990). Classification by minimum-message-length inference. In S.G. Akl, F. Fiala and W.W. Koczkodaj (eds.), *Advances in Computing and Information - ICCI'90*. Lecture Notes in Computer Science 468, (Springer-Verlag: Berlin), 72-81.

Wallace, C.S. and D.M. Boulton (1968). An information measure for classification. *Computer Journal* 11, 185-194.

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 D.L. Dowe (1994). Intrinsic classification by MML - the Snob program. *Proceedings of Seventh Australian Joint Conference on Artificial Iintelligence,* 37-44.

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 of minimum message length estimation. *Journal of the Royal Statistical Society* B 54, 195-209.