Minimum Message Length Analysis of Multiple Short Time Series

Daniel F. Schmidt and Enes Makalic

Abstract

This paper applies the Bayesian minimum message length principle to the multiple short time series problem, yielding satisfactory estimates for all model parameters as well as a test for autocorrelation. Connections with the method of conditional likelihood are also discussed.

Keywords: Minimum message length; Bayesian inference; Nuisance parameters; Approximate conditional likelihood; Gaussian autoregressive processes

1. Introduction

Consider data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)' \in \mathbb{R}^{m \times n}$ comprised of m sequences $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n})' \in \mathbb{R}^n$ generated by the following stationary first order Gaussian autoregressive model:

$$y_{i,j} = \mu_i + \varepsilon_{i,j}, \tag{1}$$
$$\varepsilon_{i,j} = \rho\varepsilon_{i,j-1} + v_{i,j}, \tag{2}$$

where $(i = 1, \ldots, m;\ j = 1, \ldots, n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)' \in \mathbb{R}^m$ are the sequence means, $\rho \in (-1, 1)$ is a common autoregressive parameter and $v_{i,j}$ denotes the innovations which are independently and identically distributed as $N(0, \tau)$. The starting point of this paper is to make inferences about the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \rho, \tau)' \in \mathbb{R}^{m+2}$ given data sampled from the model (1)–(2). The sequences are considered exchangeable, in the sense that inferences made about the model parameters should be invariant under the interchange of any pair of sequences in the matrix $\mathbf{Y}$. This model appears frequently in epidemiological and medical studies in which several measurements have been made over time on a large number of people. In this case, the autocorrelation parameter ρ is of particular interest, as it represents how well the physical quantity "tracks" over time.

Making inferences about ρ in this setting is complicated by the fact that the number of parameters grows with the number of sequences m and a straightforward application of the maximum likelihood principle leads to inconsistent estimates of both τ and ρ. A likelihood-based solution to the problem of estimating ρ in the model (1)–(2) using the method of approximate conditional likelihood was presented in [1] and shown to yield significant improvements over the standard maximum likelihood estimates. Two frequentist test procedures for the presence of autocorrelation are also discussed in [2].

A solution within the Bayesian framework of inference would be of great value. Unfortunately, with the choice of sensible priors that reflect the invariance properties required of the problem, the usual method of analysing the posterior distribution formed from the product of the prior distribution and likelihood is unsatisfactory. The posterior distribution does not concentrate probability mass around the true parameter values even as the number of sequences $m \to \infty$, and parameter estimation based on this posterior is subsequently inconsistent. This paper demonstrates that estimation based on the alternative information-theoretic Bayesian principle of minimum message length [3] leads to satisfactory estimates of all parameters $\boldsymbol{\theta}$ as well as providing a simple basis for testing for autocorrelation.

This paper has three aims: (1) to produce satisfactory point estimates for all parameters of the first order Gaussian autoregressive model, (2) to produce a suitable test for autocorrelation, and (3) demonstrate the resolution of a difficult estimation problem using the minimum message length principle.

2. Minimum Message Length

The minimum message length (MML) principle [3, 4, 5] is a Bayesian principle for inductive inference based on information theory. The essential idea behind the minimum message length principle is that compressing data is equivalent to learning structure in the data. The key measure of the quality of fit of a model to data is the length of the data after it has been compressed by the model under consideration. As the compressed data must also be decompressable, the details of the model used in the compression process must be included in the description of the data. The format of the compressed data therefore consists of two parts: the assertion, $I_{87}(\boldsymbol{\theta})$, which provides a statement of the model used to compress the data, and the detail, $I_{87}(\mathbf{y}|\boldsymbol{\theta})$, which states the data coded using the model named in the assertion. Thus, in the minimum message length framework data compression is put into a one-to-one correspondence with the traditional model selection problem.

Encoding the values of any discrete parameters used in the model is straightforward due to the direct correspondence between probability mass functions and codewords; all that is required is that a suitable prior distribution be specified for these discrete parameters. Continuous valued parameters, such as the mean of a normal distribution, are more difficult to encode, as the parameter space must be reduced from a continuum to a countable set to allow the values to be encoded. This process is at the heart of the minimum message length principle and there exists a range of techniques to obtain codelengths for distributions with continuous valued parameters [6, 5, 7]. The most commonly used approximation in the minimum message length literature is the Wallace–Freeman, or MML87, codelength approximation [5]. Let $\Theta_\gamma \in \mathbb{R}^k$ denote a parameter space of a model class indexed by $\gamma \in \Gamma$. The Wallace–Freeman approximation states that the codelength of a model $\boldsymbol{\theta} \in \Theta_\gamma$ and data $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ is

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}, \gamma) = \underbrace{-\log \pi(\boldsymbol{\theta}, \gamma) + \frac{1}{2}\log|\mathbf{J}_\gamma(\boldsymbol{\theta})| + \frac{k}{2}\log \kappa_k}_{I_{87}(\boldsymbol{\theta},\gamma)} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta}, \gamma)}_{I_{87}(\mathbf{y}|\boldsymbol{\theta},\gamma)} \tag{3}$$

where $\pi(\cdot)$ denotes a joint prior distribution over the parameter space Θ_γ and the collection of model structures under consideration $\gamma \in \Gamma$, $\mathbf{J}_\gamma(\theta)$ is the Fisher information matrix, $p(\mathbf{y}|\boldsymbol{\theta}, \gamma)$ is the likelihood function, and κ_k is the normalized second moment of an optimal quantising lattice in k-dimensions [8]. Following ([3], pp. 237), the need to determine κ_k for arbitrary dimension k is circumvented by using the approximation

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

where $\psi(\cdot)$ is the digamma function. In this paper, log is defined as the natural logarithm, and as such, all message lengths are measured in *nits* (nats), or base-e digits. The minimum message length principle advocates selecting the model $(\hat{\boldsymbol{\theta}}_{87}(\mathbf{y}, \hat{\gamma}), \hat{\gamma})$ that minimizes (3) as the most *a posteriori* likely explanation of the data given a particular choice of priors. For the sake of clarity, the explicit dependence on the model structure index γ is omitted in the rest of this paper.

The Wallace–Freeman approximation provides a unified framework for parameter estimation and model selection with two important properties: (1) the codelength is invariant under diffeomorphic transformations of the parameter space, a property not shared by other popular Bayes point estimators such as the posterior mode or posterior mean, and (2) the resultant parameter estimators have been shown to be consistent in the presence of nuisance parameters for several problems (for example, the Neyman–Scott problem [9] and factor analysis ([10, 3], pp. 297–303). While a general proof of the consistency of the MML principle in the presence of nuisance parameters does not currently exist, there have been no problems studied so far in which MML has failed to yield consistent estimates.

3. Inference of parameters in multiple short time series

3.1. Wallace–Freeman Estimates

Inference using the Wallace–Freeman estimator requires specifying a likelihood function, the corresponding Fisher information matrix and prior densities over all parameters. In the multiple short time series

setting specified by (1)–(2), the negative log-likelihood function for the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \rho, \tau)' \in \mathbb{R}^{m+2}$ is

$$-\log p(\mathbf{Y}|\boldsymbol{\theta}) = \frac{mn}{2}\log 2\pi\tau - \frac{m}{2}\log(1-\rho^2) + \frac{1}{2\tau}\sum_{i=1}^{m} T_i(\mu_i, \rho), \tag{5}$$

where

$$T_i(\mu_i, \rho) = \sum_{j=1}^{n}(y_{ij}-\mu_i)^2 + \rho^2\sum_{j=2}^{n-1}(y_{ij}-\mu_i)^2 - 2\rho\sum_{j=2}^{n}(y_{ij}-\mu_i)(y_{ij-1}-\mu_i).$$

The determinant of the Fisher information matrix for $\boldsymbol{\theta}$ is $|\mathbf{J}(\boldsymbol{\theta})| = |\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\mu})||\mathbf{J}_{\tau,\rho}(\tau,\rho)|$, where

$$|\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\mu})| = \left(\frac{(\rho-1)(n(\rho-1)-2\rho)}{\tau}\right)^m, \tag{6}$$

$$|\mathbf{J}_{\tau,\rho}(\tau,\rho)| = \frac{m^2(n-1)(n-(n-2)\rho^2)}{2\tau^2(\rho^2-1)^2}. \tag{7}$$

It remains to specify suitable prior densities over all parameters. The sequence means are location parameters and are given a uniform prior density over a compact subset of $\mathbb{R}^m$, a suitable data-driven choice for this subset being given in [11]. The innovation variance is given a conjugate scale invariant prior and the autoregression parameter an asymptotic reference prior [12]. The complete prior density over $\boldsymbol{\theta}$ is:

$$\pi(\boldsymbol{\theta}) = \pi_\mu(\mu_1,\ldots,\mu_m)\pi_\tau(\tau)\pi_\rho(\rho), \tag{8}$$

$$\pi_\mu(\mu_1,\ldots,\mu_m) \propto 1, \tag{9}$$

$$\pi_\tau(\tau) \propto \frac{1}{\tau}, \tag{10}$$

$$\pi_\rho(\rho) = \frac{1}{\pi\sqrt{1-\rho^2}}, \quad \rho \in (-1,1). \tag{11}$$

The asymptotic reference prior for the autoregressive parameter is recommended in [12] as a default Bayesian objective prior and has been shown to exhibit attractive statistical properties. The choice of uniform prior for the sequence means $\boldsymbol{\mu}$, and the scale-invariant prior for τ, are made to ensure that the resulting inferences satisfy several invariance requirements that are usually deemed to be of great importance within the social and medical sciences. With this choice of priors, inferences about all parameters remain unchanged under linear transformations of the data (translation plus scale transformations); practically, this means the inferences we make about the model parameters do not depend on the particular units in which the data are recorded, e.g., degrees Celsius as compared to degrees Fahrenheit.

Substituting (5), (6)–(7) and (8) into (3), yields the following estimates of $\boldsymbol{\mu}$ and τ:

$$\hat{\mu}_i(\rho) = \frac{\left(\sum_{j=1}^{n} y_{ij}\right) + \rho^2\left(\sum_{j=2}^{n-1} y_{ij}\right) - \rho\left(\sum_{j=2}^{n} y_{ij} + y_{ij-1}\right)}{(\rho-1)(n(\rho-1)-2\rho)}, \tag{12}$$

$$\hat{\tau}(\rho) = \frac{\sum_{i=1}^{m} T_i(\hat{\mu}_i(\rho),\rho)}{m(n-1)}. \tag{13}$$

The Wallace–Freeman estimator for $\boldsymbol{\mu}$ given ρ is equivalent to the standard maximum likelihood estimator and is unbiased. The Wallace–Freeman estimate $\hat{\rho}$ of ρ is obtained by minimising

$$I_{87}(\mathbf{Y}, \rho, \hat{\boldsymbol{\mu}}(\rho), \hat{\tau}(\rho)) = \left(\frac{mn+2}{2}\right)\log\hat{\tau}(\rho) - \left(\frac{m-1}{2}\right)\log(1-\rho^2) + \frac{m(n-1)}{2} + \frac{1}{2}\log\pi|\mathbf{J}(\hat{\boldsymbol{\mu}}(\rho), \hat{\tau}(\rho), \rho)| + \frac{mn}{2}\log 2\pi + c(m+2). \tag{14}$$

where $c(\cdot)$ is given by (4). The profile message length (14) is a continuous and differentiable function of ρ, provided $\rho \in (-1,1)$, so that numerical minimisation may be performed using a second-order procedure such as Newton–Raphson. Appendix B gives simplified estimating equations for $\hat{\rho}$ in the particular case that $n = 3$.

3.1.1. Behaviour of Wallace–Freeman Estimates

The most important property of the Wallace–Freeman estimates found by minimising (14) is that they are weakly consistent as the number of sequences $m \to \infty$.

Theorem 1. *Let ρ^* and τ^* denote the true value of ρ and τ respectively. As $m \to \infty$, the sequence of Wallace–Freeman estimates $\hat{\rho}_m$ and $\hat{\tau}_m$ obey*

$$\begin{aligned} \hat{\rho}_m &\xrightarrow{p} \rho^*, \\ \hat{\tau}_m(\hat{\rho}_m) &\xrightarrow{p} \tau^*. \end{aligned}$$

The proof is presented in Appendix A. In contrast, both the maximum likelihood estimates, and the usual maximum a posteriori (MAP) estimates found by maximising the product $p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ are inconsistent as $m \to \infty$. The inconsistency of the MAP estimate is easily established by noting that with the chosen priors (9)–(11) the term $\log \pi(\rho)\pi(\tau) = O(1)$, and is thus negligible with respect to the log-likelihood term as $m \to \infty$.

The finite sample size behaviour of the Wallace–Freeman estimator $\hat{\boldsymbol{\theta}}$ can be explored by examining the contribution of the Fisher information term to the message length. The determinant of $\mathbf{J}_\mu(\cdot)$ approaches zero as $\rho \to 1$ which acts to balance the inherent bias towards $\rho = -1$ present in the usual maximum likelihood estimates. Similarly, $|\mathbf{J}_\mu(\cdot)| \to \infty$ as $\tau \to 0$ which also acts to counter the negative bias of the maximum likelihood estimates of τ. The determinant of $\mathbf{J}_{\tau,\rho}(\cdot) \to \infty$ as $|\rho| \to 1$ which tends to shrink the autoregressive parameter towards $\rho = 0$. This is because small changes in ρ near the boundary of the parameter space lead to large changes in the dynamic behaviour of the resulting AR(1) model, and thus need to be specified with increasing precision. The simulations presented in Section 4.1 show that this property has the effect of removing a troublesome "piling-up" effect that is present in the closely related conditional likelihood estimates. The absence of "piling-up" effects in minimum message length estimates has also been observed in [13] in the case of general autoregressive moving-average (ARMA) models.

3.1.2. Jeffreys' Prior

While the particular choice of priors (9)–(11) is not critical to obtaining parameter estimates with satisfactory frequentist properties, it can be observed that using the Jeffreys' prior will render the Wallace–Freeman estimates inconsistent. The dominant terms in the Jeffreys' prior are

$$((\rho - 1)(n(\rho - 1) - 2\rho))^{\frac{m}{2}} \left(\frac{1}{\tau}\right)^{\frac{m}{2}},$$

which express a preference for $\rho \to -1$ and $\tau \to 0$, the preference growing increasingly extreme with growing m. Given the model assumptions that the sequences are exchangeable with a common τ and ρ, such a prior is clearly nonsensical; as all sequences are treated the same within the context of the model, there is no reason that we should revise our prior beliefs about ρ and τ simply by observing a greater number of sequences.

3.2. Testing for the presence of autocorrelation

An advantage of formulating a problem in the minimum message length framework is that the same measure of fit (that is, codelength) used to estimate parameters for a given model class can also be used to determine the model class itself. There is no requirement to appeal to different principles for hypothesis testing or parameter estimation. By exploiting this attractive property, we construct a simple test for the autocorrelation parameter, that is, whether $\rho \neq 0$. The test amounts to comparing the codelengths of the data compressed using a model in which ρ is fixed at zero (the no correlation model) against the correlation model with codelength given by (14). While allowing ρ to be a free parameter always results in a better fit to the data, it requires the estimation and therefore statement of one extra parameter. This leads to a hypothesis test that only includes ρ if the extra complexity of the correlation model is warranted by the data.

As two different models are being compared, it is necessary to introduce a prior distribution, $\pi_\gamma(\cdot)$, over the set $\Gamma = \{\gamma_\emptyset, \gamma_\rho\}$, where $\gamma_\emptyset$ denotes the model without correlation, and γ_ρ denotes the model with correlation. This prior distribution models the *a priori* probabilities of observing a model with or without autocorrelation. When there is no autocorrelation parameter, the model (1)–(2) reduces to the well-known Neyman–Scott problem. Using the priors (9)–(10) results in the following Wallace–Freeman codelength [9]

$$I_{87}(\mathbf{Y}, \hat{\boldsymbol{\mu}}, \hat{\tau}) = \frac{m(n-1)}{2}\left(\log \hat{\tau} + 1\right) + \frac{m}{2}\log n + \frac{1}{2}\log\frac{mn}{2} + \frac{mn}{2}\log 2\pi + c(m+1),$$

where $\hat{\boldsymbol{\mu}}$ are the usual maximum likelihood estimates, and

$$\hat{\tau} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{i,j} - \hat{\mu}_i)^2}{m(n-1)}$$

is the usual unbiased estimate of τ. Let

$$\delta = I_{87}(\mathbf{Y}, \hat{\rho}, \hat{\boldsymbol{\mu}}(\hat{\rho}), \hat{\tau}(\hat{\rho})) - I_{87}(\mathbf{Y}, \hat{\boldsymbol{\mu}}, \hat{\tau}) - \log\left(\frac{\pi_\gamma(\gamma_\rho)}{\pi_\gamma(\gamma_\emptyset)}\right)$$

denote the difference in codelengths between the no correlation and correlation model. If $\delta < 0$, the correlation model is preferred over the alternative and vice versa. Furthermore, $\exp(-\delta)$ can be directly interpreted as the posterior odds in favour of the correlation model ([3], p. 160). As the total sample size $(mn) \to \infty$, the correlation model is preferred to the no correlation model if the negative log-likelihood under the correlation model is $(1/2)\log(mn) + O(1)$ nits shorter than the negative log-likelihood for the no correlation model.

3.3. Minimum Message Length Credible Sets

In a standard Bayesian analysis, posterior credible intervals over the parameter space may be computed from the posterior density $p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Using the prior density (8), the posterior mode will be close to the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$, and as m grows, most of the posterior density will be concentrated around $(\hat{\rho}_{\mathrm{ML}}, \hat{\tau}_{\mathrm{ML}})$. Credible sets constructed in this manner are therefore expected to be unsatisfactory. An alternative construction of credible sets based on the Wallace–Freeman message length approximation is now considered. Recalling that the difference in message length between two models can be interpreted as a posterior log-odds, we propose to use the Wallace–Freeman (WF) posterior

$$w(\boldsymbol{\theta}|\mathbf{Y}) \propto \frac{p(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{|\mathbf{J}(\boldsymbol{\theta})|^{(1/2)}} \tag{15}$$

as a basis for further Bayesian analysis, where $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix for the model under consideration. The WF–posterior is equivalent to a standard posterior density formed with a special prior proportional to $\pi(\boldsymbol{\theta})|\mathbf{J}(\boldsymbol{\theta})|^{-(1/2)}$. Assuming (15) is normalizable, one may then form standard Bayesian credible sets based on the WF–posterior. This approach is essentially the same as the procedure proposed in [14] for deriving invariant credible sets and MAP estimators, but is driven by information-theoretic arguments. Given the general behaviour of the MML principle, it is expected that the WF–posterior $w(\boldsymbol{\theta}|\mathbf{Y})$ will produce satisfactory credible sets in the case where the number of parameters grows with sample size.

Consider now a regular inference problem where the number of parameters is fixed and does not grow with the total sample size denoted by t. Under suitable regularity conditions, the Fisher information matrix can be factored as $\mathbf{J}(\boldsymbol{\theta}) = t\mathbf{J}_1(\boldsymbol{\theta})$, where $\mathbf{J}_1(\theta)$ is the *per sample* Fisher information matrix. Briefly, an insight into the behaviour of the WF–posterior can be gained by applying the Laplace integral approximation to (15) which yields a Gaussian density centred on the Wallace–Freeman estimator $\hat{\boldsymbol{\theta}}$, with inverse-covariance matrix $\boldsymbol{\Sigma}^{-1}$ given by

$$\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}) = \mathbf{J}(\mathbf{Y}, \hat{\boldsymbol{\theta}}) + \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\frac{1}{2}\log|\mathbf{J}_1(\boldsymbol{\theta})| - \log\pi(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{16}$$

ρ^*	SE($\hat{\rho}$)	SE($\hat{\rho}_\lambda$)	RKL($\hat{\rho}$)	RKL($\hat{\rho}_\lambda$)	SE($\hat{\rho}$)	SE($\hat{\rho}_\lambda$)	RKL($\hat{\rho}$)	RKL($\hat{\rho}_\lambda$)
		$m = 10$				$m = 20$		
−0·90	0·01	0·01	0·24	0·65	0·00	0·00	0·14	0·55
−0·70	0·05	0·05	0·21	0·58	0·66	0·02	0·12	0·44
−0·50	0·09	0·10	0·18	0·48	0·45	0·05	0·11	0·36
−0·25	0·13	0·17	0·15	0·39	0·21	0·07	0·08	0·27
0·00	0·14	0·21	0·12	0·31	0·09	0·12	0·07	0·20
0·25	0·14	0·23	0·08	0·26	0·10	0·14	0·05	0·16
0·50	0·15	0·21	0·05	0·26	0·09	0·13	0·03	0·13
0·70	0·16	0·18	0·03	0·29	0·10	0·11	0·02	0·14
0·90	0·22	0·18	0·05	0·06	0·12	0·10	0·03	0·06

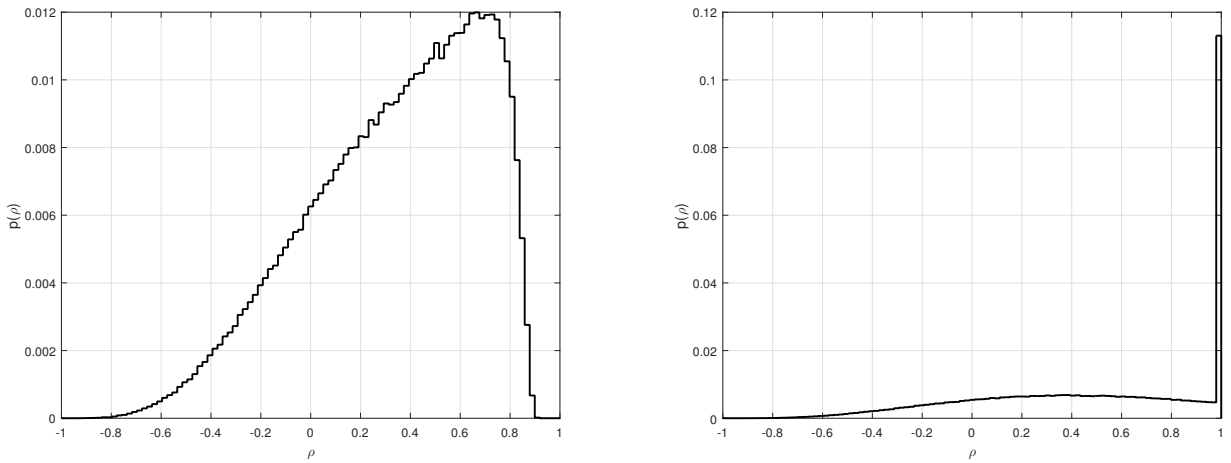
Table 1: Squared errors (SE) and relative Kullback–Leibler divergences (RKL) for the Wallace–Freeman estimates $\hat{\rho}$ and approximate conditional likelihood estimates $\hat{\rho}_\lambda$.

Here, $\mathbf{J}(\mathbf{Y}, \hat{\boldsymbol{\theta}})$ is the Hessian matrix of the negative log-likelihood and is of order $O(t)$. As the sample size $t \to \infty$, the second term in (16) is of order $O(1)$ and is swamped by the Hessian matrix, and the WF–posterior model converges on the maximum likelihood estimate, i.e., $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_{\mathrm{ML}}$; thus, increasing the sample size yields the standard posterior distribution. However, in the case of model (1)–(2) the sample size is $t = mn$, with the sequence length n fixed and the number of sequences m growing. While t samples contribute to the estimation of ρ and τ, only n samples contribute to the estimation of each μ_i, and thus the WF–posterior covariance will not converge on the regular posterior covariance even as $m \to \infty$. Furthermore, while the regular posterior mode for ρ and τ converges on the inconsistent maximum likelihood estimates, the WF–posterior mode converges on the true values of ρ and τ as $m \to \infty$, and is thus expected to give a superior measure of uncertainty for these estimates compared to the regular posterior. As in the case of the regular posterior distribution, the WF-posterior will, in general, have a non-standard form and finding credible intervals becomes problematic; for suitably regular problems, the matrix $\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$ may be used to derive approximate credible sets through the usual theory of multivariate Gaussian distributions.

4. Simulations

4.1. Estimation of ρ

The Wallace–Freeman estimates $\hat{\rho}$ of ρ were compared against the approximate conditional likelihood estimates $\hat{\rho}_\lambda$ derived in [1] and the regular maximum likelihood estimates. Due to the translation invariance of the estimates for μ given by (12), and the fact that the estimates for ρ and τ are based on the residuals $(y_{ij} - \hat{\mu}_i)$ for both the Wallace–Freeman procedure and approximate conditional likelihood, the particular choice of μ_i will have no effect on the behaviour of the estimates of ρ and τ. Therefore, all of the simulations were carried out by fixing the sequence means $\mu_i^* = 0$ $(i = 1, \ldots, m)$ and the innovation variance $\tau^* = 1$. A sequence length of $n = 3$ was chosen for the experiments; see Appendix B for simplified estimating equations for $\hat{\rho}$ and $\hat{\rho}_\lambda$ for this particular case. All tests were repeated for 10^4 iterations for all combinations of different numbers of sequences $m = \{10, 20\}$ and autocorrelation parameter $\rho^* = \{0, \pm 0{\cdot}25, \pm 0{\cdot}5, \pm 0{\cdot}7, \pm 0{\cdot}9\}$. For each combination of m and ρ^*, the mean and squared error for both the Wallace-Freeman estimates and the approximate conditional likelihood estimates were computed. The estimates were assessed in terms of squared error and Kullback–Leibler divergence for the autoregressive component of the model; the Kullback–Leibler (KL) divergence between "true" first order Gaussian autoregressive model with parameters ρ and τ



(a) Wallace–Freeman estimates

(b) Approximate conditional likelihood estimates

Figure 1: Empirical distributions of estimates of ρ where the data were generated using $\rho^* = 0.5$.

and "approximating model" with parameters $\hat{\rho}$ and $\hat{\tau}$, for a time series of length n, is given by

$$\mathrm{KL}_n(\rho, \tau || \hat{\rho}, \hat{\tau}) = \frac{1}{2}\left[\log\left(\frac{\hat{\tau}(1-\rho^2)}{\tau(1-\hat{\rho}^2)}\right) + \frac{\tau(1-\hat{\rho}^2)}{\hat{\tau}(1-\rho^2)}\right] + \left(\frac{n-1}{2}\right)\left[\frac{\tau}{\hat{\tau}} + \frac{\tau(\rho-\hat{\rho})^2}{\hat{\tau}(1-\rho^2)} + \log\left(\frac{\hat{\tau}}{\tau}\right)\right] - \frac{n}{2}. \quad (17)$$

Assessment of the approximate conditional likelihood estimator introduced using the Kullback–Leibler divergence was problematic as the solution that minimises the approximate conditonal likelihood may sometimes lie outside the stationarity region, and must be "clipped" against some upper-bound, say ρ^+. As (17) is unbounded as $|\hat{\rho} \to 1|$, the mean KL divergence for the approximate conditional likelihood estimator is primarily determined by the particular choice of upper-bound ρ^+. To circumvent this problem, we used median KL divergences, as these were essentially unaffected by different choices of the ρ^+; for all experiments we took $\rho^+ = 0{\cdot}999$. Further, to enable a comparison with the performance of the maximum likelihood estimator, the median KL divergences in Table 1 for the Wallace–Freeman and approximate conditional likelihood estimators are presented *relative* to the median KL divergences attained by the maximum likelihood estimator. The simulation results are presented in Table 1.

In terms of squared error, the $\hat{\rho}$ estimates were superior to the $\hat{\rho}_\lambda$ estimates for all but $\rho^* = 0{\cdot}9$. This can be explained by examining the empirical distributions of both estimators, an example for generating $\rho^* = 0{\cdot}5$ being shown in Figure 1. The approximate conditional likelihood estimates of ρ suffer from a "pile-up" effect [1] causing a large number of estimates to lie exactly on the upper-bound of the search space $\hat{\rho}_\lambda = \rho^+$, which clearly helps reduce squared error when $\rho^* = 0{\cdot}9$. The observed pile-up effect becomes significantly more pronounced as $\rho^* \to 1$. The empirical distribution of the Wallace–Freeman estimates exhibits no signs of a pile-up effect; this is due to the extra regularization towards $\rho = 0$ introduced by the Fisher information term $|\mathbf{J}(\rho, \tau)|$ as discussed in Section 3.1. The absence of a pile-up effect can not be attributed to the prior $\pi(\rho)$ as the reference prior for ρ is biased towards extreme values of ρ; instead, the pile-up effect absence appears to be solely due to the fact that estimates of ρ near the boundary must be stated to an increasingly higher precision, with correspondingly increased assertion length.

In terms of KL divergence, the Wallace–Freeman estimates were superior for all ρ^* examined, and for both sample sizes. Both estimators perform worse near $\rho^* = -1$, and improve in performance, relative to the maximum likelihood estimator, as $\rho^* \to 1$. In relative terms, the performance of the Wallace–Freeman estimator relative to the approximate conditional likelihood estimator is greatest near $\rho^* = -1$ and decreases as $\rho^* \to 1$. For the larger sample size, $m = 20$, both estimators improve in performance relative to the conventional maximum likelihood estimator, which conforms to expectations, given the inconsistency of the maximum likelihood estimator as $m \to \infty$.

4.2. Hypothesis Testing

The minimum message length test for autocorrelation described in Section 3.2 was compared against two test statistics developed in [2]. As per Section 4.1, 10^4 simulations for the same combinations of ρ^* and m were performed. For each iteration, the true hypothesis ($\rho^* = 0$ or $\rho^* \neq 0$) was randomly selected with equal probability and data was then generated from the chosen hypothesis. For each data sample generated, the minimum message length criterion and the two tests of Cox & Solomon were asked to nominate which of the two hypotheses (correlation or no correlation) generated the data; the prior distribution $\pi_\gamma(\gamma_\emptyset) = \pi_\gamma(\gamma_\rho) = 0{\cdot}5$ was used for all experiments.

Directly comparing the minimum message length approach against the standard hypothesis testing procedures is slightly problematic as the minimum message length test does not require specification of a Type I error; instead, the Type I and Type II error rates are automatically determined by the procedure itself, in a similar fashion to hypothesis testing via Bayes factors. To facilitate a more direct comparison, the following procedure was adopted: (i) the minimum message length criterion was used to nominate the correlation or no correlation model for each of the 10^4 experiments; (2) the observed type I error rate, $\hat{\alpha}$, of the minimum message length criterion was determined by counting the number of times the correlation model had the smaller message length in cases when the data was generated by the no correlation model; (3) the two frequentist test statistics developed in [2] were then used to nominate the correlation and no correlation model for each of the 10^4 experiments using a significance level $\alpha = \hat{\alpha}$. This matches the minimum message length and frequentist procedures on Type I error rates, ensuring a fair comparison and allowing us to compare the methods on their Type II error rates.

The resulting rates of correct hypothesis identification are presented in Table 2. The minimum message length test and the pooled Cox & Solomon test were virtually indistinguishable in all test cases, and both performed significantly better than the unpooled Cox & Solomon test. These results are encouraging given that the minimum message length test is automatically obtained simply by the comparison of codelengths. We also note that while the pooled Cox & Solomon test is difficult to extend to arbitrary sequence lengths $n > 3$, such an extension is trivial for the minimum message length test. It is also relatively straightforward to adapt the minimum message length test to the unpooled case in which each sequence has a different innovation variance. However, an interesting twist is that the codelengths of the unpooled and pooled models could further be compared to test whether all sequences shared a common innovation variance.

ρ^*	WF87	CS_V	CS_T	WF87	CS_V	CS_T
	$m = 10$			$m = 20$		
−0·90	0·96	0·85	0·96	0·98	0·96	0·98
−0·70	0·77	0·64	0·74	0·91	0·74	0·90
−0·50	0·60	0·54	0·58	0·68	0·59	0·67
−0·25	0·52	0·51	0·51	0·53	0·51	0·53
0·00	0·50	0·50	0·50	0·50	0·50	0·50
0·25	0·50	0·50	0·50	0·51	0·51	0·52
0·50	0·53	0·52	0·55	0·58	0·55	0·59
0·70	0·55	0·55	0·58	0·64	0·58	0·66
0·90	0·60	0·57	0·62	0·73	0·63	0·74

Table 2: Proportion of times Wallace–Freeman and Cox & Solomon tests correctly identify presence/absence of autocorrelation.

4.3. Credible Sets

For the model (1)–(2) with prior densities over $\boldsymbol{\theta}$ chosen as (8), the Bayesian posterior results in generally unsatisfactory credible intervals. This is because the posterior mode approaches the maximum likelihood

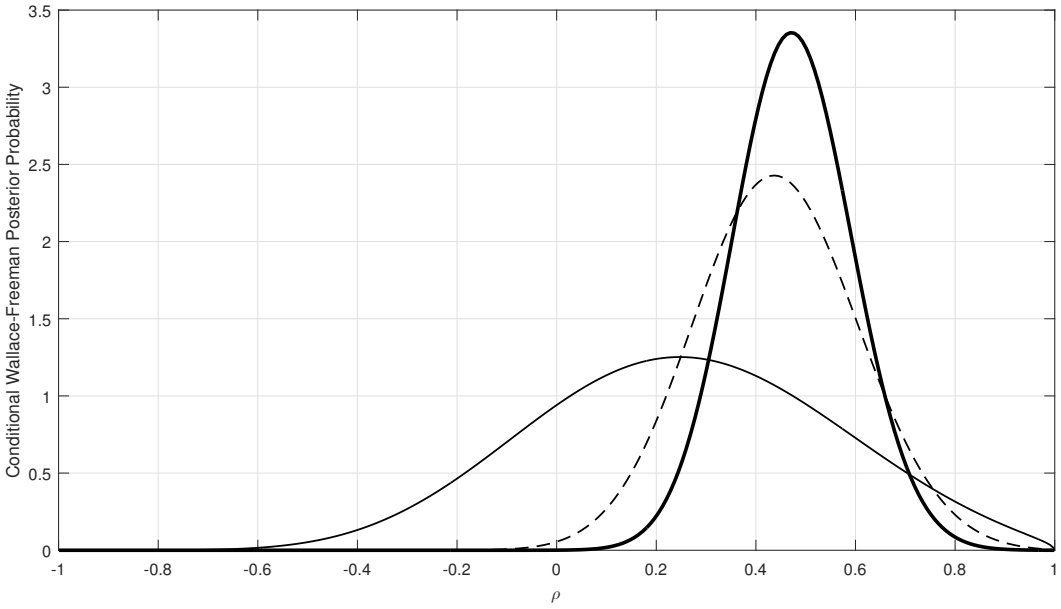


Figure 2: Average conditional Wallace–Freeman posterior probability densities for $\hat{\rho}$ generated from a model with $\rho = 0{\cdot}5$, for number of sequences $m = 10$ (light solid), $m = 50$ (dashed) and $m = 100$ (heavy solid).

estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ as $m \to \infty$ and most of the posterior mass is then concentrated away from the true parameters. In contrast, the WF–posterior (15) mode is consistent and concentrates posterior mass around the true parameter with increasing sample size. This effect is demonstrated in Figure 2 which shows the conditional WF–posterior densities (15) for ρ, averaged over 10^4 samples generated from a model with $\rho^* = 0{\cdot}5$; clearly, as m increases, the WF–posterior becomes more concentrated around the generating ρ^*. This suggests credible sets based on the WF–posterior should have satisfactory properties, at least for the model (1)–(2).

The probability of coverage for the Wallace–Freeman credible sets was then examined empirically. Due to the difficulties of normalizing and marginalizing the exact WF–posterior, the Laplace integral approximation based on (16) was used in all numerical simulations; formula for the entries of the inverse-covariance matrix $\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$ are detailed in Appendix C. Examination of the average WF–posteriors in Figure 2 suggests that this approximation should be reasonably accurate even for small m. The simulation setup was as follows: (1) the generating parameters $\boldsymbol{\theta}$ were sampled from their respective prior densities; (2) data was generated from the sampled $\boldsymbol{\theta}$ with sequence length $n = 3$, and (3) Wallace–Freeman credible sets were constructed from the WF–posterior. The number of times the resultant credible sets covered the generating ρ^* was recorded. This process was repeated 10^5 times. The results are shown in Table 3 for number of sequences $m = \{10, 50, 100, 500\}$ and desired levels of coverage $\alpha = \{0{\cdot}5, 0{\cdot}9\}$. The observed coverage attained by the Wallace–Freeman sets is in all cases approximately equal to the desired coverage. This suggests that these credible sets are a promising alternative to regular Bayesian credible sets, in particular for problems where the number of parameters grows with increasing sample size.

	$\alpha = 0.5$				$\alpha = 0.9$			
m	10	50	100	500	10	50	100	500
$\hat{\alpha}$	0·4753	0·4698	0·4658	0·4702	0·8697	0·8816	0·8840	0·8748

Table 3: Proportion of times the generating ρ^* lay inside the Wallace–Freeman credible set for coverage levels $\alpha = \{0.5, 0.9\}$

5. Comparison with Approximate Conditional Likelihood

The approximate conditional likelihood procedure can be shown to be equivalent to a restricted Wallace–Freeman estimator. In particular, the approximate conditional likelihood estimate for a parameter of interest ψ orthogonal to nuisance parameters λ can be written as

$$\hat{\psi} = \arg\min_{\psi} \left\{ I_{87}\left(y, \psi, \hat{\lambda}_{\mathrm{ML}}(\psi)\right)\right\}$$

with prior densities of (ψ, λ)

$$\begin{aligned} \pi_{\lambda}(\lambda) &\propto 1, \\ \pi_{\psi}(\psi) &\propto J_{\psi}(\psi)^{(1/2)}, \end{aligned}$$

where $\hat{\lambda}_{\mathrm{ML}}(\psi)$ is the maximum likelihood estimate of the orthogonalized parameters given ψ. Although this estimate is permissible in the sense that it produces valid codelengths, it would not be preferred in general to the full Wallace–Freeman estimates, except in the special case that the two coincide. This can only occur if the Fisher information matrix $J_{\lambda}(\lambda)$ is not a function of λ. An example in which the two estimators coincide is estimation of the variance parameter τ in the Neyman–Scott problem with prior density $\pi_{\tau}(\tau) \propto 1/\tau$.

In contrast to approximate conditional likelihood, the Wallace–Freeman estimator allows for improved estimation of all parameters without any need for sometimes difficult orthogonalization. The argument against the arbitrary choice of orthogonalization transformation can be equally applied to the arbitrary choice of subjective priors. However, the choice of priors and its interpretation is much more transparent, and it seems in general that any non-degenerate prior leads to satisfactory estimates, at least for large sample sizes. Additionally, the use of prior distributions means that the Wallace–Freeman estimates are invariant under reparameterisations of the nuisance parameters, a property not shared by the approximate conditional likelihood approach (page 236, [1]).

Furthermore, approximate conditional likelihood becomes difficult to apply if there is more than one parameter of interest. No such difficulty is present in the minimum message length approach. In the particular case of model (1)–(2), the inference of $\boldsymbol{\mu}$ poses some interesting issues. In some cases a sequence mean μ_i may realistically be the parameter of interest. It is then not entirely clear how one approaches this problem with approximate conditional likelihood. Selecting one of the μ_i as the parameter of interest results in maximum likelihood estimation of ρ, which is clearly problematic. In turn, the estimate of μ_i is expected to be unsatisfactory as it depends on ρ. Alternatively, one could first find the improved estimate $\hat{\rho}_{\lambda}$, and then perform regular maximum likelihood estimation of μ conditioned on this estimate. However, it does not seem entirely clear which procedure is preferable, and why.

Appendix A

Proof of Theorem 1. It has been shown that maximising the marginal likelihood [15] for the model (1)–(2) leads to consistent estimation of ρ [1]. To prove the consistency of the Wallace–Freeman estimate $\hat{\rho}$, it suffices to show that, asymptotically, minimising the message length (14) is equivalent to maximising the marginal likelihood. The marginal likelihood is given by

$$L_m(\rho) = (\mathbf{1}_n' \boldsymbol{\Gamma}_{\rho}^{-1} \mathbf{1}_n)^{-m/2} |\boldsymbol{\Gamma}_{\rho}|^{-m/2} \left(\sum_{i=1}^{m} \mathbf{y}_i' \mathbf{W}_{\rho} \mathbf{y} \right)^{-m(n-1)/2},$$

where $\boldsymbol{\Gamma}_{\rho}$ is the $(n \times n)$ unit-variance autocovariance matrix with entries $\Gamma_{i,j} = \mathrm{E}\left[y_{n-i} y_{n-j}\right]/\tau = \rho^{|i-j|}/(1-\rho^2)$, where y_n are samples from the autoregressive process with autoregressive parameter ρ and variance τ, $\mathbf{1}_n$ is an $(n \times 1)$ vector of ones, and

$$\mathbf{W}_{\rho} = \boldsymbol{\Gamma}_{\rho}^{-1} \left(\mathbf{I}_n - \mathbf{1}_n (\mathbf{1}_n' \boldsymbol{\Gamma}_{\rho}^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n' \boldsymbol{\Gamma}_{\rho}^{-1} \right).$$

Using the following identities

$$\begin{aligned} |\boldsymbol{\Gamma}_\rho| &= (1-\rho^2)^{-1}, \\ (\mathbf{1}_n'\boldsymbol{\Gamma}_\rho^{-1}\mathbf{1}_n)^{-m} &= \tau^{-m}|\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\mu})|^{-1}, \qquad (18) \\ \mathbf{y}_i'\mathbf{W}_\rho\mathbf{y}_i &= T_i(\hat{\mu}_i(\rho),\rho), \qquad (19) \end{aligned}$$

it is straightforward to show that

$$I_{87}(\mathbf{Y},\rho,\hat{\boldsymbol{\mu}}(\rho),\hat{\tau}(\rho)) = -\log L_m(\rho) + \frac{1}{2}\log|\mathbf{J}_{\tau,\rho}(\hat{\tau}(\rho),\rho)| - \log\pi_\tau(\hat{\tau}(\rho))\pi_\rho(\rho) + \text{const}, \qquad (20)$$

where const denotes terms independent of ρ. As $\log L_m(\rho) = O(m)$, and the second and third terms in (20) are of order $o(m)$, their contribution is negligible as $m \to \infty$ and the consistency of $\hat{\rho}$ follows. Let ρ^*, τ^* and μ_i^* denote the values of ρ, τ and μ_i from which the data was generated. To prove consistency of $\hat{\tau}(\rho)$ given by (13) we show that $\hat{\tau}(\rho^*)$ is consistent as $m \to \infty$; consistency of $\hat{\tau}(\hat{\rho})$ follows by Slutsky's theorem. The identity

$$\mathrm{E}\left[\mathbf{y}_i'\mathbf{W}_{\rho^*}\mathbf{y}_i\right] = \mathrm{Tr}\left(\boldsymbol{\Gamma}_{\rho^*}\mathbf{W}_{\rho^*}\right)\tau^* + (\mu_i^*\mathbf{1}_n)'\mathbf{W}_{\rho^*}(\mu_i^*\mathbf{1}_n) = (n-1)\tau^*,$$

may be shown by exploiting the simple tridiagonal structure of $\boldsymbol{\Gamma}_\rho^{-1}$ combined with (18). Using (19) we see that

$$\mathrm{E}\left[\hat{\tau}(\rho^*)\right] = \mathrm{E}\left[\frac{\sum_{i=1}^m T_i(\hat{\mu}_i(\rho^*),\rho^*)}{m(n-1)}\right] = \tau^*,$$

and $\mathrm{var}[\hat{\tau}(\rho)] = O(1/m)$. □

Appendix B

In this appendix the equations defining the Wallace–Freeman estimate $\hat{\rho}$ and the approximate conditional likelihood estimate $\hat{\rho}_\lambda$ are given for the special case $n = 3$. Defining

$$\begin{aligned} c_1 &= \frac{1}{m}\sum_{i=1}^m (y_{i1}^2 + y_{i2}^2 + y_{i3}^2 - y_{i1}y_{i2} - y_{i1}y_{i3} - y_{i2}y_{i3}), \\ c_2 &= \frac{1}{m}\sum_{i=1}^m (y_{i2}^2 - y_{i1}y_{i2} + y_{i1}y_{i3} - y_{i2}y_{i3}), \end{aligned}$$

the approximate conditional likelihood estimate of ρ is the root of the linear equation

$$(c_1 + c_2)\rho - c_1 + 3c_2 = 0, \qquad (21)$$

which has the solution

$$\hat{\rho}_\lambda = \min\left\{\rho^+, \frac{c_1 - 3c_2}{c_1 + c_2}\right\}. \qquad (22)$$

where $\rho^+ < 1$ is an upper-bound chosen to be close to one. The minimum clamps the estimate as the root of (21) may be greater than unity; this explains the presence of the "pile-up" effect in the approximate conditional likelihood estimates of ρ. The estimating equation (21) differs from equation 7 in [1], which is a quadratic in ρ. The consistency of (22) as $m \to \infty$ may be verified directly by noting that

$$\mathrm{E}\left[c_1\right] = \frac{\tau^*(3+\rho^*)}{1+\rho^*}, \quad \mathrm{E}\left[c_2\right] = \frac{\tau^*(1-\rho^*)}{1+\rho^*}$$

and that $\mathrm{var}[c_1]$ and $\mathrm{var}[c_2]$ are $O(1/m)$; here, ρ^* and τ^* denote the values of ρ and τ used to generate the data. The above expectation for c_2 corrects a minor typographical error present in Section 4 of [1].

The estimating equation for the Wallace–Freeman estimate $\hat{\rho}$ is

$$m(c_1+c_2)\rho^4-2(mc_1-(m-1)c_2)\rho^3-2((m+1)c_1+3(m-1)c_2)\rho^2+6((m+1)c_1-mc_2)\rho-3m(c_1-3c_2)=0, \quad (23)$$

which is a quartic polynomial in ρ. The consistency of $\hat{\rho}$ can be verified by taking the limit of (23) divided by m as $m \to \infty$ and using the above expectations to obtain

$$\rho^4 - (\rho^* + 1)\rho^3 + (\rho^* - 3)\rho^2 + 3(\rho^* + 1)\rho - 3\rho^* = 0,$$

which has roots at $\hat{\rho} = \left\{\rho^*, 1, \pm\sqrt{3}\right\}$; the codelength is infinite at $\hat{\rho} = 1$ making $\hat{\rho} = \rho^*$ the only permissible solution.

Appendix C

In this appendix we derive the approximate covariance matrix of the Wallace–Freeman estimates based on the equation (16) in Section 16. Defining

$$e_{ij} = y_{ij} - \mu_i,$$

the entries of the approximate inverse Wallace–Freeman posterior covariance matrix $\mathbf{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$ are given by

$$\begin{aligned}
\Sigma^{-1}_{\hat{\mu}_i,\hat{\mu}_i}(\hat{\boldsymbol{\theta}}) &= \frac{(\hat{\rho}-1)(n(\hat{\rho}-1)-2\hat{\rho})}{\hat{\tau}} \\
\Sigma^{-1}_{\hat{\mu}_i,\hat{\mu}_j}(\hat{\boldsymbol{\theta}}) &= \Sigma^{-1}_{\mu_i,\tau}(\hat{\boldsymbol{\theta}}) = 0 \\
\Sigma^{-1}_{\hat{\mu}_i,\hat{\rho}}(\hat{\boldsymbol{\theta}}) &= \left(\frac{1}{\tau}\right)\left(\sum_{j=2}^{n}(e_{ij}+e_{ij-1}) - 2\rho\sum_{j=2}^{n-1}e_{ij}\right) \\
\Sigma^{-1}_{\hat{\tau},\hat{\tau}}(\hat{\boldsymbol{\theta}}) &= \frac{m(n-1)}{2\hat{\tau}^2} \\
\Sigma^{-1}_{\hat{\tau},\hat{\rho}}(\hat{\boldsymbol{\theta}}) &= \left(\frac{1}{\hat{\tau}^2}\right)\sum_{i=1}^{m}\left(\sum_{j=2}^{n}e_{ij}e_{ij-1} - \rho\sum_{j=2}^{n-1}e_{ij}^2\right) \\
\Sigma^{-1}_{\hat{\rho},\hat{\rho}}(\hat{\boldsymbol{\theta}}) &= \left(\frac{1}{\hat{\tau}}\right)\sum_{i=1}^{m}\left(\sum_{i=2}^{n-1}e_{ij}^2\right) + \frac{(m-1)(\hat{\rho}^2+1)}{(\hat{\rho}^2-1)^2} + f(\hat{\rho},\hat{\tau},m,n)
\end{aligned}$$

where

$$f(\hat{\rho},\hat{\tau},m,n) = \left.\frac{\partial^2(1/2)\log|\mathbf{J}(\boldsymbol{\theta})|}{\partial\rho^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

is lengthy but straightforward to evaluate.

References

[1] A. M. Cruddas, N. Reid, D. R. Cox, A time series illustration of approximate conditional likelihood, Biometrika 76 (2) (1989) 231–237.

[2] D. R. Cox, P. J. Solomon, On testing for serial correlation in large numbers of small samples, Biometrika 75 (1) (1988) 145–148.

[3] C. S. Wallace, Statistical and Inductive Inference by Minimum Message Length, 1st Edition, Information Science and Statistics, Springer, 2005.

[4] C. S. Wallace, D. M. Boulton, An information measure for classification, Computer Journal 11 (2) (1968) 185–194. URL `http://www.allisons.org/ll/MML/Structured/1968-WB-CJ/`

[5] C. S. Wallace, P. R. Freeman, Estimation and inference by compact coding, Journal of the Royal Statistical Society (Series B) 49 (3) (1987) 240–252.

[6] C. Wallace, D. Boulton, An invariant Bayes method for point estimation, Classification Society Bulletin 3 (3) (1975) 11–34.
[7] D. F. Schmidt, A new message length formula for parameter estimation and model selection, in: Proc. 5th Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-11), 2011.
[8] J. H. Conway, N. J. A. Sloane, Sphere Packing, Lattices and Groups, 3rd Edition, Springer-Verlag, 1998.
[9] D. L. Dowe, C. S. Wallace, Resolving the Neyman-Scott problem by minimum message length, in: Proc. 28th Symposium on the interface, Vol. 28 of Computing Science and Statistics, Sydney, Australia, 1997, pp. 614–618.
[10] C. S. Wallace, P. R. Freeman, Single-factor analysis by minimum message length estimation, Journal of the Royal Statistical Society (Series B) 54 (1) (1992) 195–209.
[11] D. Schmidt, E. Makalic, MML invariant linear regression, in: The 22nd Australasian Joint Conference on Artificial Intelligence, Melbourne, Australia, 2009, pp. 312–321.
[12] J. O. Berger, R. Yang, Noninformative priors and Bayesian testing for the AR(1) model, Econometric Theory 10 (3–4) (1994) 461–482.
[13] D. F. Schmidt, Minimum message length inference of autoregressive moving average models, Ph.D. thesis, Clayton School of Information Technology, Monash University (2008).
[14] P. Druilhet, J.-M. Marin, Invariant HPD credible sets and MAP estimators, Bayesian Analsyis 2 (4) (2007) 681–692.
[15] G. Tunnicliffe-Wilson, On the use of marginal likelihood in time series model estimation, Journal of the Royal Statistical Society Series B 51 (1) (1989) 15–27.