Fast Computation of the Kullback–Leibler Divergence and Exact Fisher Information for the First-Order Moving Average Model

Enes Makalic and Daniel F. Schmidt, *Member*, *IEEE*

***Abstract*—In this note expressions are derived that allow computation of the Kullback–Leibler divergence between two first-order Gaussian moving average models in $O_n(1)$ time as the sample size $n \to \infty$. These expressions can also be used to evaluate the exact Fisher information matrix in $O_n(1)$ time, and provide a basis for an asymptotic expression of the Kullback–Leibler divergence.**

***Index Terms*—Moving Average Models, Kullback–Leibler divergence, Fisher Information**

I. Introduction

Consider the first-order moving average, MA(1), explanation for a sequence of data $\mathbf{y} = (y_1, \ldots, y_n)'$,

$$y_i = v_i - \phi v_{i-1}, \quad (i = 1, \ldots, n) \tag{1}$$

where $\phi \in (-1, 1)$ is the moving average parameter, and $v_i \sim N(0, \tau)$ are independently and identically distributed normal innovations. Let $\boldsymbol{\theta} = (\phi, \tau)$ denote the MA(1) parameters; the likelihood of $\mathbf{y}$ given $\boldsymbol{\theta}$ is the multivariate normal distribution

$$-\log f(\mathbf{y}|\theta) = \frac{n}{2}\log 2\pi\tau + \frac{1}{2}\log|\boldsymbol{\Gamma}(\phi)| + \frac{1}{2\tau}\mathbf{y}'\boldsymbol{\Gamma}^{-1}(\phi)\mathbf{y}, \tag{2}$$

where $\boldsymbol{\Gamma}(\phi)$ is the $(n \times n)$ unit variance autocovariance matrix, with entries

$$\Gamma_{i,j}(\phi) = \begin{cases} \gamma_0 = 1 + \phi^2 & \text{for } |i-j| = 0 \\ \gamma_1 = -\phi & \text{for } |i-j| = 1 \\ \gamma_{|i-j|} = 0 & \text{for } |i-j| \geq 2 \end{cases}. \tag{3}$$

Let $\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})$ denote the Kullback–Leibler divergence [1] between the "true" model $\boldsymbol{\theta}_* = (\phi_*, \tau_*)$ and the "approximating" model $\boldsymbol{\theta} = (\phi, \tau)$. The Kullback–Leibler divergence is an important, parameterisation-free loss function with information theoretic interpretations that also forms the basis for several other loss functions, such as Kullback's symmetric divergence and Bernardo's intrinsic loss. For multivariate normal distributions, of which model (1) is a special case, the Kullback–Leibler divergence, $\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})$, is given by

$$\frac{n}{2}\log\frac{\tau}{\tau_*} + \frac{1}{2}\log\frac{|\boldsymbol{\Gamma}(\phi)|}{|\boldsymbol{\Gamma}(\phi_*)|} + \left(\frac{\tau_*}{2\tau}\right)\mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Gamma}^{-1}(\phi)\right) - \frac{n}{2}. \tag{4}$$

Direct evaluation of (4) has computational complexity of order $O_n(n^3)$; the purpose of this note is to derive expressions to compute the Kullback–Leibler divergence between two MA(1) models in $O_n(1)$ time. As a corollary of this, one may also find the exact Fisher information for the MA(1) model in $O_n(1)$ time.

Enes Makalic and Daniel F. Schmidt are with the Centre for MEGA Epidemiology, The University of Melbourne, Carlton, VIC 3053, Australia

Email: {emakalic,dschmidt}@unimelb.edu.au

II. Fast Computation of the Kullback–Leibler Divergence

The $O_n(1)$ time complexity expressions for the $\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})$ are now derived. We first note that the determinant of $\boldsymbol{\Gamma}(\phi)$ reduces to [2]

$$|\boldsymbol{\Gamma}(\phi)| = \sum_{j=0}^{n}\phi^{2j} = \frac{\phi^{2n+2} - 1}{\phi^2 - 1}, \tag{5}$$

so that the first two terms of (4) are

$$\frac{n}{2}\log\frac{\tau}{\tau_*} + \frac{1}{2}\log\left(\frac{(\phi^{2n+2} - 1)(\phi_*^2 - 1)}{(\phi_*^{2n+2} - 1)(\phi^2 - 1)}\right). \tag{6}$$

Letting $\mathbf{I}_n$ denote the $(n \times n)$ identity matrix, the third term of (4) may be rewritten using the alternative expression for $\boldsymbol{\Gamma}^{-1}(\phi)$ developed in [3] as

$$\boldsymbol{\Gamma}^{-1}(\phi) = \frac{\boldsymbol{\Omega}(\mathbf{I}_n + \boldsymbol{\Lambda})}{(1 - \phi^2)}, \tag{7}$$

where $\Omega_{i,j} = \phi^{|i-j|}$ is the autocorrelation matrix of the auxiliary autoregressive process associated with ϕ, $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}, 0_{(n\times(n-2))}, \tilde{\mathbf{I}}_n\boldsymbol{\lambda})'$, $\tilde{\mathbf{I}}_n$ is the $(n \times n)$ permutation matrix that reverses the order of the elements of $\boldsymbol{\lambda}$, and

$$\lambda_j = \frac{\phi^{j+1}\left(1 - \phi^{2(n-j+1)}\right)}{\phi^{2n+2} - 1}, \quad (j = 1, \ldots, n). \tag{8}$$

Using (7) yields

$$\left(1 - \phi^2\right)\mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Gamma}^{-1}(\phi)\right) = \mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}\right) + \mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}\boldsymbol{\Lambda}\right). \tag{9}$$

The first trace on the right hand side of (9) may be evaluated by noting that the diagonal of the product $\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}$ contains only two distinct elements. The first and last diagonal entries are $(\gamma_0 + \gamma_1\phi)$, and the $(n-2)$ diagonal entries in between are all $(\gamma_0 + 2\gamma_1\phi)$, yielding

$$\mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}\right) = (\phi_*^2 - 2\phi_*\phi + 1)n + 2\phi\phi_*. \tag{10}$$

It remains to evaluate the final trace in (9). First, we note that $\mathrm{tr}\left(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}\boldsymbol{\Lambda}\right) = \mathrm{tr}(\boldsymbol{\Lambda}(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega}))$. Due to the special structure of $\boldsymbol{\Lambda}$, the product $\boldsymbol{\Lambda}(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega})$ has identical top-left and bottom-right elements, with all the other diagonal entries set to zero. Thus it suffices to compute the inner product of $\boldsymbol{\lambda}$

with the first column of $(\boldsymbol{\Gamma}(\phi_*)\boldsymbol{\Omega})$, say $\mathbf{p}$. This column vector has entries

$$p_i = \begin{cases} \gamma_0 + \gamma_1\phi & \text{for } i = 1 \\ (\gamma_1 + \gamma_0\phi + \gamma_1\phi^2)\phi^{i-2} & \text{for } 2 \leq i \leq (n-1) \\ \gamma_1\phi^{n-2} + \gamma_0\phi^{n-1} & \text{for } i = n \end{cases}, \tag{11}$$

so that the inner product $(\boldsymbol{\lambda}'\mathbf{p})$ consists of three distinct terms, say $b_1 = \lambda_1 p_1$, $b_2 = \lambda_n p_n$ and $b_3 = \sum_{i=2}^{n-1} \lambda_i p_i$, given by

$$b_1 = \frac{(1 + \phi_*^2 - \phi\phi_*)\phi^2(1 - \phi^{2n})}{\phi^{2n+2} - 1},$$
$$b_2 = -\frac{(\phi^2 - 1)(\phi^{2n}(1 + \phi_*^2) - \phi^{2n-1}\phi_*)}{\phi^{2n+2} - 1},$$
$$b_3 = \frac{(1 - \phi\phi_*)(\phi_* - \phi)(\phi^{2n}((n-2)(\phi^4 - \phi^2) - 1) + \phi^4)}{\phi(\phi^2 - 1)(\phi^{2n+2} - 1)},$$

where $\lim_{\phi \to 0} \{b_1 + b_2 + b_3\} = 0$. The Kullback–Leibler divergence is then given by

$$\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta}) = \frac{n}{2}\log\frac{\tau}{\tau_*} + \frac{1}{2}\log\left(\frac{(\phi^{2n+2} - 1)(\phi_*^2 - 1)}{(\phi_*^{2n+2} - 1)(\phi^2 - 1)}\right) - \frac{n}{2} + \left(\frac{\tau_*}{2\tau}\right)\left(\frac{(\phi_*^2 - 2\phi_*\phi + 1)n + 2\phi\phi_* + 2(b_1 + b_2 + b_3)}{1 - \phi^2}\right) \tag{12}$$

which takes $O_n(1)$ time to evaluate.

Remark 1 (Boundary Cases). The expression (12), and the representation (9) that it is based on break down when $|\phi| = 1$, i.e. the approximating model lies on the invertibility boundary resulting in infinite Kullback–Leibler divergence. This deficiency may be overcome by handling these special cases using the limit of (12) as $|\phi| \to 1$, yielding

$$\lim_{|\phi|\to 1}\{\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})\} = \frac{n}{2}\log\frac{\tau}{\tau_*} + \frac{1}{2}\log\left(\frac{(n+1)\phi_*^2 - n - 1}{\phi_*^{2n+2} - 1}\right) + \left(\frac{n\tau_*}{12\tau}\right)\left((n+2)\phi_s^2 - (2n-2)\phi_s + n + 2\right) - \frac{n}{2} \tag{13}$$

with $\phi_s = \mathrm{sgn}(\phi)\phi_*$.

Remark 2 (Entropy). The entropy, $H(\boldsymbol{\theta}_*)$, of an MA(1) model may be found by using (5), yielding

$$H(\boldsymbol{\theta}_*) = \frac{n}{2}\log 2\pi + \frac{n}{2}\log\tau_* + \frac{1}{2}\log\left(\frac{\phi^{2n+2} - 1}{\phi^2 - 1}\right) + \frac{n}{2},$$

which may be evaluated in $O_n(1)$ time.

III. The Fisher Information Matrix

Using (12) we can derive an $O_n(1)$ time complexity expression for the exact Fisher information matrix. Under suitable regularity conditions, the Fisher information matrix is given by

$$\mathbf{J}(\boldsymbol{\theta}^*) = \left[\left.\frac{\partial\boldsymbol{\theta}^2\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]. \tag{14}$$

Plugging (12) into (14) yields the following entries for $\mathbf{J}(\cdot)$

$$J_{\phi,\phi}(\boldsymbol{\theta}) = \frac{\phi^{4n}c_1 + \phi^{2n}c_2 - (n+3)\phi^2 + n - 1}{(\phi^2 - 1)^2(\phi^{2n+2} - 1)^2}, \tag{15}$$

$$J_{\phi,\tau}(\boldsymbol{\theta}) = \frac{\phi^{2n}(\phi^3 - \phi)n - \phi^{2n+1} + \phi}{\tau(\phi^{2n}(\phi^4 - \phi^2) - \phi^2 + 1)}, \tag{16}$$

$$J_{\tau,\tau}(\boldsymbol{\theta}) = \frac{n}{2\tau^2}, \tag{17}$$

with $J_{\tau,\phi}(\boldsymbol{\theta}) = J_{\phi,\tau}(\boldsymbol{\theta})$, and

$$c_1 = 2\phi^6 n - (3n+5)\phi^4 + (n+1)\phi^2,$$
$$c_2 = (2n^2 + 3n + 5)\phi^4 - (4n^2 + 6n - 2)\phi^2 + 2n^2 + 3n + 1.$$

Using (15)–(17) one may also compute the exact unnormalised Jeffrey's prior, $\pi(\theta) \propto |\mathbf{J}(\boldsymbol{\theta})|^{1/2}$ in $O_n(1)$ time.

IV. Large Sample Behaviour

The large-sample behaviour of the Kullback–Leibler divergence (12) is now briefly examined. One can break the Kullback–Leibler divergence into two terms

$$\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta}) = \Delta_n(\boldsymbol{\theta}_*||\boldsymbol{\theta}) + \Delta_1(\boldsymbol{\theta}_*||\boldsymbol{\theta})$$

where $\Delta_n(\cdot) = O_n(n)$ and $\Delta_1(\cdot) = O_n(1)$. The asymptotic Kullback–Leibler divergence may then be found by taking limits of the per sample Kullback–Leibler divergence, $\Delta(\boldsymbol{\theta}_*||\boldsymbol{\theta})/n$, as $n \to \infty$, yielding

$$\frac{1}{2}\log\frac{\tau}{\tau_*} + \left(\frac{\tau_*}{2\tau}\right)\left(\frac{\phi_*^2 - 2\phi_*\phi + 1}{1 - \phi^2}\right) - \frac{1}{2}, \tag{18}$$

with $\Delta_n(\boldsymbol{\theta}_*||\boldsymbol{\theta})$ given by n times (18). The difference between $\Delta_n(\boldsymbol{\theta}_*||\boldsymbol{\theta})$ and the exact Kullback–Leibler divergence, given by $\Delta_1(\boldsymbol{\theta}_*||\boldsymbol{\theta})$, while being $O_n(1)$, can be very large for values of ϕ close to the invertibility boundary. In fact, we note that (18) tends to infinity as $|\phi| \to 1$ while the exact Kullback–Leibler divergence is always finite, at least for finite n, as seen from (13). It is interesting to compare (18) to the asymptotic per sample Kullback–Leibler divergence for an AR(1) model with autoregressive parameter equal to ϕ_* given by

$$\frac{1}{2}\log\frac{\tau}{\tau_*} + \left(\frac{\tau_*}{2\tau}\right)\left(1 + \frac{(\phi - \phi_*)^2}{1 - \phi_*^2}\right) - \frac{1}{2}. \tag{19}$$

An interestingly duality is that in the case of the MA(1) model, the asymptotic Kullback–Leibler divergence is unbounded as $|\phi| \to 1$ and the exact expression remains bounded, while in the case of the AR(1) model the exact expression is unbounded as $|\phi| \to 1$ and the asymptotic approximation is bounded. Although (18) does not coincide with (19), the asymptotic Fisher information matrices derived from these expressions do coincide, agreeing with the results in [4].

References

[1] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[2] W. Dixon, "Further contributions to the problem of serial correlation," *The Annals of Mathematical Statistics*, vol. 15, no. 2, pp. 119–144, 1944.

[3] J. N. Haddad, "On the closed form of the likelihood function of the first order moving average model," *Biometrika*, vol. 82, no. 1, pp. 232–234, 1995.

[4] P. Whittle, "The analysis of multiple stationary time series," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 15, no. 1, pp. 125–139, 1953.