Estimating the Order of an Autoregressive Model using Normalized Maximum Likelihood

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

***Abstract*—This paper examines the estimation of the order of an autoregressive model using the Minimum Description Length principle. A closed form for an approximation of the parametric complexity of the autoregressive model class is derived by exploiting a relationship between coefficients and partial autocorrelations. The parametric complexity over the complete parameter space is found to diverge. A model selection criterion is subsequently derived by bounding the parameter space, and simulations suggest that it compares well against standard autoregressive order selection techniques in terms of correct order identification and prediction error.**

***Index Terms*—Gaussian Autoregressive Processes, Minimum Description Length, Normalized Maximum Likelihood, Maximum Likelihood Estimation**

I. Introduction

CONSIDER the p_*-th order autoregressive, AR(p_*), explanation for a sequence $\mathbf{y}^n = (y_1, \ldots, y_n) \in \mathbb{R}^n$

$$y_n + \sum_{i=1}^{p_*} \phi_i y_{n-i} = v_n \tag{1}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{p_*})$ are the autoregressive coefficients and $v_n \sim N(0, \tau)$ are the independently and identically distributed normal innovations. Such models find extensive use in disciplines as widely varied as engineering, finance and meteorology, and a standard problem is the selection of a suitable order, p, based purely on the n observed datapoints $\mathbf{y}^n$. Perhaps the most common methods for obtaining an estimate of p_* are based on the minimization of penalized likelihood scores such as the Akaike Information Criterion (AIC) [1], the Bayesian Information Criterion (BIC) [2] (sometimes mistakenly called the "MDL" criterion) or the symmetric Kullback Information Criterion (KIC) [3]. Let $f_p(\cdot)$ and $\hat{\boldsymbol{\phi}}(\mathbf{y}^n) = \arg\max_{\boldsymbol{\phi}} \{f_p(\mathbf{y}^n|\boldsymbol{\phi})\}$ denote the likelihood function and the maximum likelihood (ML) estimator for the AR(p) model respectively. These approaches can be unified by the Generalized Information Criterion (GIC)

$$\mathrm{GIC}(p) = -2 \log f_p\left(\mathbf{y}^n|\hat{\boldsymbol{\phi}}(\mathbf{y}^n)\right) + p\alpha \tag{2}$$

where $\alpha = 2$ for AIC, $\alpha = \log n$ for BIC and $\alpha = 3$ for KIC. All these techniques are based on asymptotic arguments and act by penalizing the likelihood with a penalty that, for a given sample size n, is proportional to the number of fitted parameters. Finite sample criteria that are not based on the same asymptotic assumptions as AIC and KIC, such as the corrected AIC [4] (AIC$_c$) and KIC [5] (KIC$_c$) also exist, and these are designed to work well even when the ratio $n/p > 1$ is close to unity. Although both of these small sample criteria are designed with regular linear models in mind, they may be applied in the autoregressive case by appealing to the asymptotic distribution of the maximum likelihood estimates for autoregressive models. Finite sample considerations have also lead to the Combined Information Criterion (CIC) [6] which is designed explicitly for autoregressive model order selection.

An alternative approach to estimating p_* is based on the relationship between compression and model selection; these ideas are embodied in the Minimum Message Length [7] and Minimum Description Length (MDL) [8], [9] principles. The essential idea is that the model class that most compresses the data should be chosen as the best explanation. This paper examines the problem of selecting a suitable order p for an autoregression using the MDL principle, and in particular, the Normalized Maximum Likelihood (NML) model. The NML criterion is the negative logarithm of the special NML density, which for the autoregressive model class is given by

$$f_{\mathrm{NML}}(\mathbf{y}^n|p) = \frac{f_p(\mathbf{y}^n|\hat{\boldsymbol{\phi}}(\mathbf{y}^n))}{\int_{\mathbf{x}^n \in \mathbb{R}^n} f_p(\mathbf{x}^n|\hat{\boldsymbol{\phi}}(\mathbf{x}^n)) d\mathbf{x}^n} \tag{3}$$

with $\hat{\boldsymbol{\phi}}(\cdot)$ the maximum likelihood estimator. In MDL terminology the negative logarithm of (3) is referred to as the *stochastic complexity* of the data $\mathbf{y}^n$, and the denominator of (3) is usually referred to as the *parametric complexity*. The parametric complexity has two important interpretations: first, the logarithm of the parametric complexity is the regret attained by the minimax codebook for the dataspace $\mathcal{Y}^n$ with respect to the ideal codelength given by the negative logarithm of the numerator in (3), and second, it may be viewed as counting the number of *distinguishable distributions* that are contained within the model class of interest [10]. This latter interpretation is closely related to the requirement that continuous parameter spaces be discretized when constructing codes, and offers a direct measure of the richness of a model class. The NML model has several strong theoretical properties that make it highly attractive as a model selection tool [11] and may be interpreted as providing a "universal sufficient statistic" representation of the data. In many cases the normalizing integral in (3) is extremely difficult to compute, and one may instead resort to the approximation of Rissanen [12] which is given by

$$\begin{aligned} -\log f_{\mathrm{NML}}(\mathbf{y}^n|p) &= -\log f_p\left(\mathbf{y}^n|\hat{\boldsymbol{\phi}}(\mathbf{y}^n)\right) + \frac{p}{2}\log\frac{n}{2\pi} \\ &\quad + \log \int_{\Phi} \sqrt{|\mathbf{J}(\boldsymbol{\phi})|}\, d\boldsymbol{\phi} + o(1) \end{aligned} \tag{4}$$

Daniel F. Schmidt and Enes Makalic are with the Centre for MEGA Epidemiology, The University of Melbourne, Carlton, VIC 3053, Australia
Email: {dschmidt,emakalic}@unimelb.edu.au

where Φ is the parameter space and $\mathbf{J}(\cdot)$ is the *per sample* Fisher information matrix. It is clear that for large n the Fisher information term in (4) is of order $O(1)$ with respect to n and the NML criterion differs little from BIC; however, the logarithm of the integral term in (4) contains a large amount of extra structural information about a model class that can significantly alter behaviour for small to moderate sample sizes.

To date, the only published attempt to evaluate (4) in the case of AR(p) models was undertaken in [13]. Due to the difficulty in evaluating the integral directly, a numerical approach based on quasi–random sampling was used. To overcome the problem of sampling from the highly irregular polytope that defines the stationarity region in coefficient space, the AR(p) model was reparameterized in terms of the roots of its characteristic polynomial; this provides a simple stationarity region (a hyper-box) and uniform sampling is straightforward. However, there appears to be two problems with this work. The first, as is shown in the sequel, is that the parametric complexity for an AR(p) model with $p > 1$ is infinite. This is in contrast to the finite values of the parametric complexity obtained in [13]; we believe this is likely due to the samples being drawn from a compact subset of the parameter space, i.e., sampling poles with a magnitude less than $1 - \epsilon$, for some $\epsilon > 0$, and that a different choice of ϵ would result in a different parametric complexity. The second problem is that the parametric complexities are estimated using only a single pole configuration for each order p (either $p/2$ complex roots for p even, or $\lfloor p/2 \rfloor$ complex roots and one real root for p odd); however, there are $\lfloor p/2 \rfloor + 1$ possible combinations of real and imaginary poles that define the complete AR(p) model space, and [13] does not check the pole configuration of the resulting maximum likelihood estimates. Restriction to a single pole configuration, irrespective of the pole configuration of the maximum likelihood estimates, means that the resulting parametric complexity may not even include the model which has been fitted to the data.

This paper makes two contributions: (a) by exploiting links between autoregressive coefficients and partial autocorrelations it derives a closed form expression for the integral term in (4), showing that it diverges for all $p > 1$, and (b) using the resulting expressions and a particular bounding of the parameter space an MDL criterion for autoregressive model selection is derived. This new NML procedure offers an interesting alternative consistent procedure to the more commonly used, and relatively conservative BIC criterion, that performs well if the observed data is believed to contain a moderate to large amount of structure.

II. Likelihood Function

The exact likelihood function for an AR(p) model is given by a multivariate normal distribution with an $(n \times n)$ covariance matrix. Following [14] this may be more compactly written as

$$-\log f_p(\mathbf{y}^n|\boldsymbol{\phi},\tau) = \frac{n}{2}\log(2\pi\tau) + \frac{1}{2}\log|\boldsymbol{\Gamma}(\boldsymbol{\phi})| + \frac{1}{2\tau}\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \tag{5}$$

where $\boldsymbol{\beta} = (1, \boldsymbol{\phi})' \in \mathbb{R}^{p+1}$, $\mathbf{D} \in \mathbb{R}^{(p+1\times p+1)}$ is a matrix with the entries $D_{i,j} = \sum_{k=0}^{n-i-j+1} y_{k+i}y_{k+j}$, and $\boldsymbol{\Gamma}(\boldsymbol{\phi}) \in \mathbb{R}^{(p\times p)}$ is the unit-variance process autocovariance matrix with entries $\Gamma_{i,j}(\boldsymbol{\phi}) = (1/\tau)\mathrm{E}\left[y_{n-i}y_{n-j}\right]$ [15].

A. Maximum Likelihood Estimation

The maximum likelihood estimate for τ, given some estimates $\boldsymbol{\phi}$, is

$$\hat{\tau}(\mathbf{y}^n) = \frac{1}{n}\boldsymbol{\beta}'\mathbf{D}\boldsymbol{\beta} \tag{6}$$

The maximum likelihood estimates of $\boldsymbol{\phi}$ must be found by numerical optimisation; the primary advantage of the likelihood representation (5) is that once $\mathbf{D}$ has been formed, subsequent evaluations of (5) require only $O(p^2)$ operations. This paper shall be restricted to stationary autoregressive processes; the stationarity restrictions in $\boldsymbol{\phi}$-space are defined by a complex polytope. Alternatively, one may parameterize the AR(p) model in terms of p partial autocorrelations $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$, in which the stationarity region becomes the interior of a hypercube centered on the origin. There exists a one-to-one, continuous mapping between coefficients and partial autocorrelations given by the Levinson–Durbin recurrence relations [16]. Using these the likelihood function (5) may be expressed in terms of the partial autocorrelations; denote this by $-\log f_p(\mathbf{y}^n|\boldsymbol{\rho},\tau)$. An important simplification of working with partial autocorrelations is that the evaluation of the half log-determinant term in (5) reduces to [17]

$$\frac{1}{2}\log|\boldsymbol{\Gamma}(\boldsymbol{\phi})| = -\frac{1}{2}\sum_{j=1}^{p} j\log(1-\rho_j^2) \tag{7}$$

For the numerical search it is desirable to have the gradient and the Hessian of $-\log f_p(\mathbf{y}^n|\boldsymbol{\rho},\tau)$ with respect to $\boldsymbol{\rho}$. Formulae to compute these quantities based on the Levinson-Durbin recurrence relations are presented in Appendix A.

III. Parametric Complexity of Autoregressive Models

An important property of the autoregressive model class is that despite the fact that the data $\mathbf{y}^n$ are serially correlated, the effects of this dependency are of order $O(1)$. For suitably large n the asymptotic *per sample* information matrix is given by [15]

$$\mathbf{J}(\boldsymbol{\phi},\tau) = \lim_{n\to\infty}\left\{\frac{\mathbf{J}_n(\boldsymbol{\phi},\tau)}{n}\right\} = \begin{pmatrix} \boldsymbol{\Gamma}(\boldsymbol{\phi}) & 0 \\ 0 & \dfrac{1}{2\tau^2} \end{pmatrix}, \tag{8}$$

where $\mathbf{J}_n(\boldsymbol{\phi},\tau)$ is the exact information matrix in coefficient space for n datapoints, and $\boldsymbol{\Gamma}(\boldsymbol{\phi})$ is the $(p \times p)$ unit-variance process autocovariance matrix. Given that the noise variance τ does not appear in $\boldsymbol{\Gamma}(\boldsymbol{\phi})$ the $O(1)$ component of the stochastic complexity in (4) may be written as

$$\int_{\Theta}\sqrt{|\mathbf{J}(\boldsymbol{\theta})|}d\boldsymbol{\theta} = \int_{\tau_0}^{\tau_1}\left(\frac{1}{2\tau^2}\right)^{\frac{1}{2}}d\tau\int_{\Phi}\sqrt{|\boldsymbol{\Gamma}(\boldsymbol{\phi})|}\,d\boldsymbol{\phi},$$

where Φ is the set of stationary models in coefficient space, i.e., those models for which all roots of the characteristic polynomial $1+\sum_{i=1}^{p}\phi_i z^{-i}$ lie within the unit circle. It is clear that

as in [13] the noise variance parameter $\tau \in (\tau_0, \tau_1)$ may be safely ignored when computing the stochastic complexity (4); including the variance simply scales the parametric complexity for all model orders by a constant determined by the arbitrary range of integration (τ_0, τ_1), and thus has no effect on model selection.

It remains to compute the integral

$$\int_{\Phi} \sqrt{|\mathbf{\Gamma}(\boldsymbol{\phi})|}\, d\boldsymbol{\phi}. \tag{9}$$

The highly complex shape of the admissible parameter region Φ coupled with the difficulties of evaluating the determinant in coefficient space results in an integral that is extremely difficult to evaluate analytically; hence the attempt in [13] to overcome this problem by numerical evaluation of the integral in root space, the deficiencies of which have been previously discussed. An alternative approach undertaken in this paper is to work in partial autocorrelation space; as in Section II let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$ denote the vector of partial autocorrelations corresponding to the coefficients $\boldsymbol{\phi}$. Recalling that

$$|\mathbf{\Gamma}(\boldsymbol{\phi})| = \prod_{j=1}^{p} \frac{1}{(1-\rho_j^2)^j},$$

it is clear that the integral (9) is considerably easier to deal with in partial autocorrelation space. The determinant of the Jacobian of the transformation from partial autocorrelation space to coefficient space is given in [18] as

$$|\mathbf{T}(\boldsymbol{\rho})| = \prod_{j=1}^{p} (1-\rho_j)^{\lfloor j/2 \rfloor} (1+\rho_j)^{\lfloor (j-1)/2 \rfloor}.$$

Using the standard rule for transforming integrals it is possible to rewrite (9) as an integral in partial autocorrelation space

$$\int_{\Phi} \sqrt{|\mathbf{\Gamma}(\boldsymbol{\phi})|}\, d\boldsymbol{\phi} = \int_{P} |\mathbf{T}(\boldsymbol{\rho})| \prod_{j=1}^{p} \left(\frac{1}{(1-\rho_j^2)^{j/2}} \right) d\boldsymbol{\rho}, \tag{10}$$

where $P = \{\boldsymbol{\rho} : |\rho_j| < 1, j = 1, \ldots, p\}$, i.e., the interior of a hypercube centred at the origin and of side length two. Due to the independence of the components of $\boldsymbol{\rho}$ in (10), the integral may instead be evaluated efficiently as a product of integrals. After expanding and simplifying, (10) becomes

$$\prod_{j=1}^{p} \int_{P} (1-\rho_j)^{(\lfloor j/2 \rfloor - j/2)} (1+\rho_j)^{(\lfloor (j-1)/2 \rfloor - j/2)} d\rho_j. \tag{11}$$

The indefinite integrals are given by

$$\arcsin(\rho_j) \qquad \text{for } j \text{ odd}, \tag{12}$$

$$\log(\rho_j + 1) \qquad \text{for } j \text{ even}. \tag{13}$$

It is clear from (13) that the definite integral over the region $(-1, 1)^p$ diverges for $p > 1$. Following the suggestion in [12] the next section develops an NML criterion by restricting the region of integration to a compact subset of $(-1, 1)^p$.

IV. An NML Criterion for Autoregressive Order Selection

Using the results of the previous section, an NML criterion for autoregressive model order selection can be defined. We use the term "an NML criterion", rather than "the criterion" because the problem of infinite parametric complexity requires a bounding of the parameter space. There is at present no single agreed best way for this to be done. We choose a method that is both simple in terms of computation, as well as arguably "simple" in a complexity sense, in that it requires only a single hyperparameter. The restricted parameter set is formed so that $|\rho_j| \leq \xi < 1$ for all j. This restriction results in integration over the hypercube $[-\xi, \xi]^p$, which preserves the efficient evaluation of the complexity as a product of integrals. Given a ξ, the definite integrals of (11) from $[-\xi, \xi]$ are

$$2 \arcsin(\xi) \qquad \text{for } j \text{ odd}, \tag{14}$$

$$2 \operatorname{arctanh}(\xi) \qquad \text{for } j \text{ even}. \tag{15}$$

Using (14) and (15) the total stochastic complexity for an AR(p) model, conditioned on ξ and $p \geq 1$, is given by

$$\begin{aligned} \mathrm{SC}(\mathbf{y}^n | \xi, p) = &- \log f_p\left(\mathbf{y}^n | \hat{\boldsymbol{\rho}}(\mathbf{y}^n), \hat{\tau}(\mathbf{y}^n)\right) + \frac{p}{2} \log \frac{n}{2\pi} \\ &+ \lceil p/2 \rceil \log \arcsin(\xi) + \lfloor p/2 \rfloor \log \operatorname{arctanh}(\xi) \\ &+ p \log 2 + o(1), \end{aligned} \tag{16}$$

where $\hat{\boldsymbol{\rho}}(\mathbf{y}^n)$ are the maximum likelihood estimates of the partial autocorrelations. These may be found numerically using the equations for the gradient and Hessian presented in Appendix A and a suitable constrained gradient descent or second order search procedure, such as Newton–Raphson [19]. It remains to choose a suitable ξ; a simple, fixed choice such as $\xi = 1 - \epsilon$ is arbitrary, with the choice of ϵ having a crucial effect on the criterion. As an alternative, we choose ξ to solve

$$\hat{\xi}(\mathbf{y}^n) = \underset{\xi \in (0,1)}{\arg\min} \left\{ \mathrm{SC}(\mathbf{y}^n | \xi, p) \right\} \tag{17}$$

subject to the constraint that $\hat{\boldsymbol{\rho}}(\mathbf{y}^n) \in (-\xi, \xi)^p$, i.e., that the compact set contains the maximum likelihood estimates of $\boldsymbol{\rho}$. This choice of ξ minimises the codelength for the string $\mathbf{y}^n$, and the solution to (17) is $\hat{\xi}(\mathbf{y}^n) = ||\hat{\boldsymbol{\rho}}(\mathbf{y}^n)||_\infty$. Of course, for the stochastic complexity to represent a true codelength the hyperparameter ξ must also be transmitted.

Ideally, as in [20], the NML procedure would be applied to $\mathrm{SC}(\mathbf{y}^n | \xi, p)$, treating ξ as a regular continuous parameter. Unfortunately, this does not seem possible and we instead opt to use a simple two-part code as in [21] to transmit ξ; this inflates the total stochastic complexity by $(1/2) \log n$, which is constant across all model orders $p \geq 1$. Finally, one must also construct a code for the chosen model order; given the nested nature of autoregressive models a uniform code over some range $\{0, \ldots, q\}$ with q taken to be appropriately large (in practice $q \ll n/2$ would be considered reasonable). This introduces a further term of $\log(q+1)$, which is constant across model orders and thus also has no effect on the criterion. The final stochastic complexity, ignoring the irrelevant $\log(q+1)$ term, is then given by

$$\mathrm{SC}(\mathbf{y}^n, \hat{\xi}(\mathbf{y}^n), p) = \mathrm{SC}(\mathbf{y}^n | \hat{\xi}(\mathbf{y}^n), p) + \frac{1}{2} \log n \tag{18}$$

in the case that $p \geq 1$. An important special case that is not handled by (18) is $p = 0$, i.e., no serial correlation in the data. This can be handled by coding the data $\mathbf{y}^n$ using a zero-mean normal distribution with a free variance parameter τ. Using the maximum likelihood estimate $\hat{\tau}(\mathbf{y}^n) = \mathbf{y}'\mathbf{y}/n$ yields the stochastic complexity, up to constants, of

$$\mathrm{SC}(\mathbf{y}^n, p = 0) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\frac{\mathbf{y}'\mathbf{y}}{n}\right) + \frac{n}{2}. \quad (19)$$

There is no requirement to transmit the hyperparameter ξ so the extra $(1/2)\log n$ term in (18) may be omitted. Given an observed data sequence $\mathbf{y}^n$, the "optimal" order is found by solving

$$\hat{p}(\mathbf{y}^n) = \arg\min_p \left\{ \mathrm{SC}(\mathbf{y}^n, \hat{\xi}(\mathbf{y}^n), p) \right\},$$

where $p \in \{0, \ldots, q\}$ ranges over the orders under consideration.

A. Consistency of $\hat{p}(\mathbf{y}^n)$

For a fixed region of integration (i.e., a fixed ξ), it is clear that the estimate $\hat{p}(\mathbf{y}^n)$ is strongly consistent. However, in (18) the bound $\hat{\xi}(\mathbf{y}^n)$ is a function of $\mathbf{y}^n$, and the integral term in the parametric complexity can potentially grow arbitrarily large. Consistency of $\hat{p}(\mathbf{y}^n)$ can be established by the following theorem.

Theorem 1: Let the data $\mathbf{y}^n$ be generated by an AR(p_*) model with partial autocorrelations $\boldsymbol{\rho}_*$, and let the maximum candidate order $q \geq p_*$ be independent of n. Then, as $n \to \infty$

$$\int_{\Phi\left(\hat{\xi}(\mathbf{y}^n)\right)} \sqrt{|\boldsymbol{\Gamma}(\boldsymbol{\phi})|}\, d\boldsymbol{\phi} \to c_{\boldsymbol{\rho}_*} \quad \text{a.s.},$$

where $\Phi(\xi)$ is the set of coefficients which map to partial autocorrelations $\boldsymbol{\rho}$ satisfying $||\boldsymbol{\rho}||_\infty \leq \xi$, and $c_{\boldsymbol{\rho}_*}$ is a constant depending only on $\boldsymbol{\rho}_*$.

Proof: First consider the case when $p \geq p_*$; by the consistency of the maximum likelihood estimator sequence $\hat{\boldsymbol{\rho}}(\mathbf{y}^n)$ [22] it follows that $\hat{\xi}(\mathbf{y}^n) \to ||\boldsymbol{\rho}_*||_\infty$ almost surely as $n \to \infty$. In the case that $p < p_*$, the maximum likelihood estimator sequence converges on the parameter values that minimise the Kullback–Leibler divergence from $\boldsymbol{\rho}_*$, and it therefore follows that $\hat{\xi}(\mathbf{y}^n) \to ||\arg\min_{\boldsymbol{\rho}} \{\mathrm{KL}(\boldsymbol{\rho}_*||\boldsymbol{\rho})\}||_\infty$ almost surely as $n \to \infty$. □

Theorem 1 shows that the integral in (3) converges almost surely to an $O(1)$ constant term as $n \to \infty$, and is therefore dominated by the $(p/2)\log n$ term. The stochastic complexity criterion (18) is thus asymptotically equivalent to BIC, and it follows that $\hat{p}(\mathbf{y}^n)$ is a strongly consistent estimator of p_* [23].

V. Discussion

A. Exact Information Matrix

The stochastic complexity criterion (18) was derived using the asymptotic Fisher information matrix for the autoregressive model class. For finite samples the exact Fisher information $\mathbf{J}_n(\boldsymbol{\phi})$ is (for almost all parameter values) different from n times the asymptotic matrix, and it is conceivable that using the exact matrix may render the parametric complexity finite. However, by noting that $|\mathbf{J}_n(\boldsymbol{\phi})/n| \geq |\mathbf{J}(\boldsymbol{\phi})|$ for all $\boldsymbol{\phi}$, it is clear that replacing the asymptotic Fisher information matrix in (9) by $\mathbf{J}_n(\boldsymbol{\phi})/n$ does not result in a finite integral. It is possible that the use of the exact information matrix may result in a criterion with improved small sample order selection properties; the main obstacle to this is that the determinant of the exact information matrix does not admit a simple expression in terms of partial autocorrelations as does the asymptotic information matrix.

B. Behaviour of the Penalty Terms

The penalty term in (18) exhibits an interesting "adaptive" behaviour in a similar fashion to the MDL linear regression criterion [20]. It has been noted that the MDL regression criterion "adjusts" its penalty depending on the nature of the generating model; if the generating model is comprised of many coefficients with small effects (i.e., low signal to noise ratio), the penalty terms remain small. However, the presence of a strong signal component (i.e., high signal to noise ratio) significantly increases the magnitude of the penalty terms, and the resulting criterion is subsequently less likely to overfit [9]. The signal-to-noise ratio of an AR(p) model with partial autocorrelations $\boldsymbol{\rho}$ is the ratio of signal variance to noise variance and is given by [24]

$$\mathrm{SNR}(\boldsymbol{\rho}) = \frac{\mathrm{E}\left[y_i^2\right] - \mathrm{E}\left[v_i^2\right]}{\mathrm{E}\left[v_i^2\right]} = \frac{\tau(\gamma_0 - 1)}{\tau} = \prod_{j=1}^{p} \frac{1}{(1-\rho_j^2)} - 1 \quad (20)$$

where $\gamma_0 = (1/\tau)\mathrm{E}\left[y_i^2\right]$ is the unit-variance zero order autocovariance. It is clear from (20) that the signal-to-noise ratio increases with increasing $|\rho_j|$ (< 1). The penalty terms in (18) grow with increasing ξ, which itself grows as the maximum partial autocorrelation grows, and thus the stochastic complexity behaves in a similar fashion to the case of MDL linear regression; the presence of many small partial autocorrelations (with resulting low contribution to the signal component of the time series) means that $\hat{\xi}(\mathbf{y}^n)$ will be small, and the resulting penalty terms will also be small, leading to a decreased probability of underfitting. However, the introduction of a large partial autocorrelation, which corresponds to the presence of a strong signal component in the time series (and correspondingly increases the signal-to-noise ratio of the fitted model), results in a significantly increased penalty term just as in the MDL linear regression criterion and a subsequently reduced probability of overfitting.

C. Moving Average Models

The work in [13] presents an interesting argument exploiting the asymptotic equivalence of the Fisher information for the autoregressive and moving average (MA) models to apply their NML criterion to model selection for both AR and MA models. The result of Section III shows that the parametric complexity of the autoregressive model class is in general

infinite, and thus if one were to use the asymptotic equivalence of AR and MA models the same problem of infinite parametric complexity would arise. In contrast, the moving average parameter space is compact, and the exact Fisher information for the pure moving average model class is finite over the complete moving average parameter space, even on the boundaries. Thus, the parametric complexity of moving average models is finite for any finite n and does not require any bounding of the moving average parameter space. Unfortunately, the time complexity for computing the exact information matrix for anything but the MA(1) model [25] is of order $O(n^2)$ [15], and determining the parametric complexity would seem to be possible only through a computationally expensive numerical approach.

D. Jeffreys Prior

The unnormalized uninformative Jeffreys prior [26] is given by $\pi_J(\boldsymbol{\phi}) \propto \sqrt{|\mathbf{J}(\boldsymbol{\phi})|}$, and hence the $O(1)$ term in the parametric complexity is the normalization term required to make the Jeffreys prior a proper probability density. An interesting observation is that for autoregressive models of order $p > 1$, the Jeffreys prior over the complete parameter space is improper.

VI. Simulations

A simulation study was undertaken to investigate the performance of the NML criterion. Two performance measures were chosen: (1) the frequency with which each criterion selected the "true" generating model order, and (2) the model error obtained by the resulting estimated model. The model error is defined as the normalised expected squared prediction error obtained by using an estimated model, say $\hat{\boldsymbol{\phi}}$, to predict data arising from the true model, say $\boldsymbol{\phi}_*$ and is given by

$$\mathrm{ME}(\hat{\boldsymbol{\phi}}, \boldsymbol{\phi}_*) = (\boldsymbol{\phi}_* - \hat{\boldsymbol{\phi}})' \boldsymbol{\Gamma}(\boldsymbol{\phi}_*)(\boldsymbol{\phi}_* - \hat{\boldsymbol{\phi}})/\gamma_0 \qquad (21)$$

where $\gamma_0 = \mathrm{E}\left[y_t^2\right]$ and y_t is generated by $\boldsymbol{\phi}_*$ with unit variance. The simulation compared the NML criterion given by (18) against the AIC$_c$, KIC$_c$, BIC, and CIC criteria. These are well known methods that are widely used throughout the literature to select autoregressive orders, and thus offer a good yardstick against which to assess the performance of the new NML criterion.

Additionally, the NML criterion was also compared against the new Sequentially Normalised Least Squares (SNLS) criterion [27] that is inspired by the Sequentially Normalised Maximum Likelihood procedure [28]. This criterion has been derived for regular linear models, and to apply it in the autoregressive setting a lag-matrix of the data $\mathbf{y}^n$ must be formed to act as covariates. To fit an AR(p) via this linear regression method the first p data points must be sacrificed to build the lag-matrix; furthermore, the SNLS procedure itself requires sacrificing the first p data point-covariate pairs, so that the SNLS criterion for an AR(p) model is based on $(n - 2p)$ data points rather than the full n. To ensure that the SNLS scores for the different orders of AR models are comparable, the data used was based on the maximum order q; thus, the SNLS scores were calculated using the last $(n-2q)$ datapoints for all values of p. The model errors were calculated using maximum likelihood parameter estimates for the model order chosen by SNLS, rather than the least-squares parameter estimates produced implicitly by the SNLS procedure, as these are not gauranteed to be stationary. This makes the SNLS model errors commensurate with the other criteria under consideration.

The simulations were undertaken using generating models of ten true orders $p_* = \{1, \ldots, 10\}$. At each of the $1,000$ iterations, the partial autocorrelations for the true generating models were randomly sampled so that the $r^2 = 1 - \tau/\gamma_0$ values (i.e., proportion of variance explained by the "signal" component of the series) were uniformly distributed on $(0, 1)$. This may be done following the procedure described in [29], and ensures that the simulations adequately covered a range of generating models with different signal-to-noise ratios. Given that the noise variance parameter τ has no effect on the signal-to-noise ratio, it was set to $\tau = 1$ in all tests. Using the resulting generating model, data sequences of length $n = \{50, 100, 200, 400, 800\}$ were generated. From these data sequences, the maximum likelihood estimates for AR(p) models were computed, with $p = \{0, \ldots, 10\}$ (i.e., $q = 10$), and all model selection criteria were required to nominate a suitable model order. This was then recorded along with the resulting model error obtained by the chosen models. The exception was the CIC criterion, which used the Burg estimates rather than the maximum likelihood estimates, as discussed in [30]. The results, in terms of number of times each criterion selected the true generating order, as well as the model errors (with standard errors given in parenthesis), are presented in Table I and II.

The results suggest that the NML criterion is competitive in terms of both correct model order identifications as well as prediction error. The order selection results show three basic trends. When the generating model is low order ($p_* \leq 2$), the more conservative BIC criterion performs well as there is little structure to discover in the data. In contrast, when the generating model is high order ($p_* \geq 8$) the AIC$_c$ performs the best; this is believed to be due to the tendency of AIC to select complex models, particularly if there is a large amount of structure in the data. Given that the maximum possible order considered is $q = 10$, selecting a high order model will obviously lead to good performance when p_* is close to q. In between these two extremes the NML criterion performs very strongly, and is competitive with AIC even when p_* is large. It is important to note that even though AIC$_c$ or KIC$_c$ outperform our NML criterion in terms of correct order identification for some combinations of (p_*, n) they are not consistent criteria and so will necessarily perform worse on this measure as n grows large.

In direct comparison to the consistent BIC and SNLS criteria, which the NML criterion is in some sense designed to replace, it is clear that for $p_* \geq 3$ the NML criterion uniformly attains a higher proportion of correct order identifications, the improvements increasing for larger p_*. The NML criterion also generally outperforms BIC and SNLS in terms of prediction errors for $p_* > 3$. The poor performance of the SNLS criterion is attributed to the use of linear regression models

p_*	n	Model selection criteria					
		AIC_c	BIC	KIC_c	CIC	SNLS	NML
1	50	745	**869**	839	793	836	776
	100	742	**917**	870	851	894	869
	200	734	**943**	869	857	929	905
	400	714	**971**	873	860	965	948
	800	698	**971**	876	870	966	955
2	50	578	605	**614**	569	584	570
	100	616	**703**	687	671	682	686
	200	647	**794**	756	743	786	771
	400	690	**843**	804	803	841	833
	800	693	893	828	829	**899**	884
3	50	470	441	458	440	422	**476**
	100	535	547	562	552	543	**567**
	200	594	647	650	640	660	**667**
	400	623	731	715	714	728	**741**
	800	651	781	761	759	787	**791**
4	50	357	319	340	331	311	**359**
	100	462	455	482	477	478	**490**
	200	514	516	545	**546**	530	541
	400	557	594	601	596	604	**628**
	800	551	659	657	655	674	**681**
5	50	261	205	216	227	206	**266**
	100	**351**	296	344	338	306	334
	200	417	387	**447**	444	406	426
	400	454	472	492	491	473	**495**
	800	486	522	564	563	539	**566**
6	50	**241**	180	197	211	186	232
	100	**331**	275	310	311	274	316
	200	387	354	**401**	395	373	397
	400	449	434	**483**	479	438	465
	800	469	486	**512**	512	494	508
7	50	197	142	146	167	155	**204**
	100	**284**	230	271	274	236	276
	200	336	290	323	327	292	**340**
	400	394	358	**407**	405	360	406
	800	449	404	460	**468**	411	441
8	50	154	112	116	144	96	**171**
	100	**239**	177	202	221	168	212
	200	**321**	240	289	297	236	281
	400	**376**	298	365	364	300	344
	800	407	350	409	**410**	350	378
9	50	139	92	87	129	93	**159**
	100	**223**	163	190	189	152	209
	200	**301**	205	269	272	200	256
	400	**348**	256	331	331	244	299
	800	**381**	320	370	368	312	341
10	50	114	81	79	107	58	**143**
	100	**229**	134	169	197	114	197
	200	**333**	195	270	285	175	256
	400	**405**	266	351	359	242	321
	800	**478**	315	415	419	287	365

TABLE I
NUMBER OF TIMES TRUE ORDER WAS SELECTED BY MODEL SELECTION CRITERIA FOR SIMULATED DATA

p_*	n	Model selection criteria					
		AIC_c	BIC	KIC_c	CIC	SNLS	NML
1	50	0.065 (0.004)	**0.046** (0.005)	0.049 (0.005)	0.147 (0.041)	0.048 (0.005)	0.068 (0.005)
	100	0.032 (0.002)	**0.018** (0.001)	0.021 (0.001)	0.035 (0.004)	0.020 (0.001)	0.023 (0.001)
	200	0.016 (0.001)	**0.008** (0.000)	0.010 (0.001)	0.013 (0.001)	0.008 (0.000)	0.010 (0.001)
	400	0.008 (0.000)	**0.003** (0.000)	0.005 (0.000)	0.006 (0.000)	0.003 (0.000)	0.003 (0.000)
	800	0.004 (0.000)	**0.002** (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)	0.002 (0.000)
2	50	0.088 (0.004)	0.078 (0.004)	**0.075** (0.004)	0.133 (0.015)	0.088 (0.007)	0.093 (0.004)
	100	0.043 (0.002)	**0.034** (0.001)	0.035 (0.001)	0.049 (0.003)	0.035 (0.001)	0.038 (0.002)
	200	0.021 (0.001)	0.015 (0.001)	0.016 (0.001)	0.021 (0.002)	**0.015** (0.001)	0.016 (0.001)
	400	0.010 (0.000)	0.007 (0.000)	0.008 (0.000)	0.008 (0.000)	**0.007** (0.000)	0.007 (0.000)
	800	0.005 (0.000)	0.003 (0.000)	0.004 (0.000)	0.004 (0.000)	**0.003** (0.000)	0.003 (0.000)
3	50	0.147 (0.022)	0.147 (0.022)	**0.142** (0.022)	0.448 (0.161)	0.153 (0.023)	0.161 (0.023)
	100	0.055 (0.002)	0.051 (0.002)	**0.049** (0.002)	0.097 (0.023)	0.051 (0.002)	0.052 (0.002)
	200	0.027 (0.001)	0.023 (0.001)	0.023 (0.001)	0.039 (0.009)	**0.022** (0.001)	0.023 (0.001)
	400	0.013 (0.001)	0.010 (0.001)	0.011 (0.001)	0.040 (0.028)	0.010 (0.001)	**0.010** (0.001)
	800	0.007 (0.001)	0.005 (0.001)	0.006 (0.001)	0.013 (0.007)	**0.005** (0.001)	0.005 (0.001)
4	50	**0.153** (0.010)	0.163 (0.010)	0.156 (0.010)	0.304 (0.047)	0.181 (0.014)	0.161 (0.010)
	100	0.064 (0.002)	0.061 (0.002)	**0.059** (0.002)	0.093 (0.010)	0.059 (0.002)	0.060 (0.002)
	200	0.032 (0.002)	0.031 (0.002)	**0.029** (0.002)	0.037 (0.002)	0.030 (0.002)	0.030 (0.002)
	400	0.015 (0.001)	0.014 (0.000)	0.014 (0.000)	0.018 (0.003)	0.013 (0.000)	**0.013** (0.000)
	800	0.007 (0.000)	0.006 (0.000)	0.006 (0.000)	0.007 (0.000)	0.006 (0.000)	**0.006** (0.000)
5	50	**0.169** (0.007)	0.193 (0.008)	0.184 (0.008)	0.800 (0.313)	0.210 (0.011)	0.182 (0.008)
	100	0.078 (0.003)	0.081 (0.003)	**0.074** (0.003)	0.251 (0.102)	0.079 (0.003)	0.077 (0.003)
	200	0.036 (0.001)	0.036 (0.001)	**0.033** (0.001)	0.067 (0.012)	0.035 (0.001)	0.034 (0.001)
	400	0.018 (0.001)	0.017 (0.001)	0.017 (0.001)	0.021 (0.003)	0.017 (0.001)	**0.016** (0.001)
	800	0.009 (0.000)	0.008 (0.000)	0.008 (0.000)	0.009 (0.001)	0.008 (0.000)	**0.007** (0.000)
6	50	**0.218** (0.015)	0.255 (0.016)	0.238 (0.015)	1.803 (1.299)	0.289 (0.025)	0.229 (0.015)
	100	**0.093** (0.004)	0.105 (0.004)	0.095 (0.004)	0.210 (0.048)	0.104 (0.004)	0.096 (0.004)
	200	0.040 (0.001)	0.043 (0.001)	**0.038** (0.001)	0.058 (0.009)	0.040 (0.001)	0.039 (0.001)
	400	0.019 (0.001)	0.018 (0.001)	**0.017** (0.001)	0.030 (0.009)	0.018 (0.001)	0.017 (0.001)
	800	0.010 (0.001)	0.010 (0.001)	**0.009** (0.001)	0.012 (0.002)	0.009 (0.001)	0.009 (0.001)
7	50	**0.248** (0.013)	0.287 (0.013)	0.276 (0.013)	2.732 (1.648)	0.287 (0.013)	0.262 (0.013)
	100	**0.100** (0.003)	0.115 (0.004)	0.102 (0.003)	0.258 (0.060)	0.108 (0.003)	0.102 (0.003)
	200	0.047 (0.002)	0.052 (0.002)	0.045 (0.001)	0.160 (0.065)	0.050 (0.002)	**0.045** (0.001)
	400	0.022 (0.001)	0.023 (0.001)	0.021 (0.001)	0.035 (0.007)	0.023 (0.001)	**0.021** (0.001)
	800	0.011 (0.000)	0.011 (0.000)	**0.010** (0.000)	0.012 (0.001)	0.011 (0.000)	0.010 (0.000)
8	50	**0.267** (0.016)	0.330 (0.037)	0.319 (0.037)	0.637 (0.134)	0.355 (0.039)	0.285 (0.017)
	100	**0.113** (0.006)	0.131 (0.007)	0.118 (0.007)	0.512 (0.315)	0.133 (0.007)	0.119 (0.007)
	200	**0.048** (0.001)	0.056 (0.002)	0.049 (0.001)	0.269 (0.205)	0.055 (0.002)	0.050 (0.001)
	400	0.023 (0.001)	0.026 (0.001)	**0.023** (0.001)	0.035 (0.010)	0.026 (0.001)	0.023 (0.001)
	800	0.012 (0.000)	0.013 (0.000)	**0.011** (0.000)	0.014 (0.002)	0.013 (0.000)	0.011 (0.000)
9	50	**0.313** (0.016)	0.389 (0.019)	0.379 (0.019)	0.578 (0.090)	0.421 (0.024)	0.333 (0.017)
	100	**0.121** (0.004)	0.146 (0.005)	0.128 (0.004)	0.216 (0.047)	0.151 (0.006)	0.127 (0.004)
	200	**0.053** (0.002)	0.064 (0.002)	0.055 (0.002)	0.071 (0.005)	0.064 (0.002)	0.056 (0.002)
	400	**0.025** (0.001)	0.031 (0.001)	0.026 (0.001)	0.034 (0.006)	0.031 (0.001)	0.027 (0.001)
	800	0.015 (0.002)	0.016 (0.002)	**0.014** (0.002)	0.015 (0.002)	0.016 (0.002)	0.015 (0.002)
10	50	**0.324** (0.012)	0.373 (0.015)	0.369 (0.014)	0.630 (0.066)	0.408 (0.017)	0.332 (0.012)
	100	**0.142** (0.006)	0.176 (0.007)	0.153 (0.006)	0.274 (0.061)	0.182 (0.008)	0.149 (0.006)
	200	**0.064** (0.003)	0.078 (0.004)	0.067 (0.004)	0.090 (0.011)	0.080 (0.004)	0.068 (0.004)
	400	**0.028** (0.001)	0.036 (0.001)	0.029 (0.001)	0.056 (0.024)	0.036 (0.001)	0.031 (0.001)
	800	**0.013** (0.000)	0.016 (0.000)	0.013 (0.000)	0.016 (0.001)	0.017 (0.000)	0.014 (0.000)

TABLE II
MODEL ERRORS OBTAINED BY MODEL SELECTION CRITERIA FOR SIMULATED DATA

Measure	n	Model selection criteria					
		AIC_c	BIC	KIC_c	CIC	SNLS	NML
Average Selected Order	25	4.38	4.74	3.65	4.72	5.00	4.98
	50	6.37	5.99	5.98	6.35	5.62	6.16
	100	7.00	6.34	6.75	6.86	6.18	6.45
Squared Prediction Error	25	152.10 (2.64)	158.39 (3.09)	167.21 (2.96)	**148.93** (2.51)	159.51 (2.85)	156.68 (3.02)
	50	**110.65** (0.76)	110.59 (0.72)	110.03 (0.67)	110.77 (0.72)	115.83 (1.16)	110.85 (0.76)
	100	**102.07** (0.56)	103.41 (0.46)	102.83 (0.54)	102.42 (0.60)	103.52 (0.45)	103.26 (0.48)

TABLE III
AVERAGE SELECTED ORDER AND SQUARED PREDICTION ERRORS FOR THE REAL DATASET

to approximate an autoregressive processes, as well as the requirement to sacrifice the first $2q = 20$ datapoints. The NML criterion is usually less conservative than BIC if the generating process has low signal to noise ratio (see Section V-B), and thus performs worse when $p_* = 1$; a conservative criterion having an obvious advantage when there is little structure in the data to discover. The general conclusion that can be drawn from these experiments is that if one wishes to use an asymptotically consistent criterion, then NML offers an interesting alternative to the BIC criterion, particularly if one believes there to be a moderate to large amount of structure in the data. This is frequently the case when analysing real data.

The somewhat poorer performance of the NML criterion for $p_* = 1$ in general, and for $(p_* = 1, n = 50)$ in particular, is possibly due to the use of the asymptotic rather than exact information matrix to compute the parametric complexity. The stochastic approximation as approximated by (18) seems to lead to insufficiently large penalties when fitting an AR(1) model in comparison to the penalties for $p > 1$, i.e., the *per parameter* penalty for $p > 1$ appears to be significantly lower than for $p = 1$, which manifests itself as a tendency to overfit when the generating model is $p_* = 1$. This is in contrast to BIC in which the *per parameter* penalty is fixed at $(1/2) \log n$. Of course, it should be remembered that the criterion (18) is just *one* possible NML criterion; it is distinctly possible that a different bounding of the parameter space could lead to a criterion with superior performance. Given the strong performance of NML for $p_* > 1$ we believe that improving the approximation for the case that $p_* = 1$ would be a worthwhile avenue of future research.

A. Real Data

Finally, the performance of the NML criterion was assessed on a real time series. The dataset chosen was from the physics domain and consisted of 320 measurements of the force exerted on a cylinder placed inside a tank of water [31]. The dataset was divided into overlapping blocks of length $n = \{25, 50, 100\}$ starting at sample $q + 1$, each block beginning one sample after the previous, i.e., the first block was at sample $q + 1$, the second block at sample $q + 2$, etc. Autoregressive models of order $p = \{0, \ldots, 10\}$ (i.e., $q = 10$) were fitted to the blocks, and all six criteria (AIC_c, BIC, KIC_c, CIC, SNLS and NML) were required to select one of the fitted models. The resulting models were then validated by computing the one-step-ahead squared prediction error on *all* datapoints in the dataset except those that were used to fit the model (i.e., the current block). The results are summarised in Table III.

The NML criterion obtains lower prediction errors than SNLS for all sample sizes, and outperforms BIC for $n = 25$ and $n = 100$. The good performance of CIC is somewhat surprising in light of the simulation results, but is corroborated by reports of strong performance found in the literature [6]. The NML criterion is competitive with AIC_c and KIC_c, outperforming KIC_c for $n = 25$.

APPENDIX A: GRADIENT AND HESSIAN

For completeness equations for the gradient and Hessian of the negative log-likelihood for an AR(p) model in partial autocorrelation space are presented; similar expressions may be used to compute the exact and asymptotic Fisher information matrix [32] if required. Given a vector $\boldsymbol{\rho} \in (-1, 1)^p$ of partial autocorrelations the corresponding coefficients $\boldsymbol{\phi}$ are found by

$$\begin{aligned} \boldsymbol{\phi}^{(1)} &= -\rho_1 \\ \boldsymbol{\phi}^{(k)} &= \left(\boldsymbol{\phi}^{(k-1)} - \rho_k \tilde{I} \boldsymbol{\phi}^{(k-1)}, -\rho_k\right), \quad (k > 1) \end{aligned}$$

where $\boldsymbol{\phi}^{(k)}$ denotes the k-dimensional coefficient vector at iteration k of the recurrence relations, $\tilde{I}$ is the permutation matrix that reverses the order of the elements of a vector, and $\boldsymbol{\phi} \equiv \boldsymbol{\phi}^{(p)}$. The components of the gradient and Hessian for the last term in (5) depend on the vectors $\partial\boldsymbol{\phi}/\partial\rho_j$. These vectors may themselves be formed through the recurrence relations

$$\begin{aligned} \partial\boldsymbol{\phi}^{(1)}/\partial\rho_j &= \begin{cases} -\rho_1 & \text{for } j > 1 \\ -1 & \text{for } j = 1 \end{cases} \\ \partial\boldsymbol{\phi}^{(k)}/\partial\rho_j &= \begin{cases} \left(\boldsymbol{\phi}^{(k-1)} - \rho_k \tilde{I} \boldsymbol{\phi}^{(k-1)}, -\rho_k\right) & \text{for } j < k \\ \left(-\tilde{I} \boldsymbol{\phi}^{(k-1)}, -1\right) & \text{for } j = k \\ \left(\boldsymbol{\phi}^{(k-1)} - \rho_k \tilde{I} \boldsymbol{\phi}^{(k-1)}, 0\right) & \text{for } j > k \end{cases} \end{aligned}$$

so that $\partial\boldsymbol{\beta}/\partial\rho_j = (0, \partial\boldsymbol{\phi}^{(p)}/\partial\rho_j)$, and we may form the matrix $\partial\boldsymbol{\beta}/\partial\boldsymbol{\rho}' = (\partial\boldsymbol{\beta}/\partial\rho_1, \ldots, \partial\boldsymbol{\beta}/\partial\rho_p)$. The half log-determinant of $\boldsymbol{\Gamma}(\boldsymbol{\phi})$ in (5) is succinctly given by [17]

$$\frac{1}{2} \log |\boldsymbol{\Gamma}(\boldsymbol{\phi})| = -\frac{1}{2} \sum_{j=1}^{p} j \log(1 - \rho_j^2) \tag{22}$$

The gradient is then given by

$$-\frac{\partial \log f_p(\mathbf{y}^n|\boldsymbol{\rho},\tau)}{\partial \boldsymbol{\rho}'} = \mathbf{b}' + \boldsymbol{\beta}' \left(\frac{\mathbf{D}}{\tau}\right)\left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\rho}'}\right)$$

where $\mathbf{b} \in \mathbb{R}^p$ is a vector with entries $b_j = j\rho_j/(1-\rho_j^2)$ which are the partial derivatives of (22) with respect to ρ_j, respectively. The Hessian is given by

$$-\frac{\partial^2 \log f_p(\mathbf{y}^n|\boldsymbol{\rho},\tau)}{\partial \boldsymbol{\rho}\partial \boldsymbol{\rho}'} = \mathbf{A} + \left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\rho}'}\right)' \left(\frac{\mathbf{D}}{\tau}\right)\left(\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\rho}'}\right) \quad (23)$$

where $\mathbf{A} \in \mathbb{R}^{(p\times p)}$ is a diagonal matrix with non-zero elements

$$A_{j,j} = \frac{j(\rho_j^2+1)}{(\rho_j^2-1)^2}$$

which are the second partial derivatives of (22) with respect to ρ_j.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.

[2] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[3] J. E. Cavanaugh, "A large-sample model selection criterion based on Kullback's symmetric divergence," *Statistics & Probability Letters*, vol. 42, no. 4, pp. 333–343, 1999.

[4] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, June 1989.

[5] A.-K. Seghouane and M. Bekara, "A small sample model selection criterion based on Kullback's symmetric divergence," *IEEE Transactions on Signal Processing*, vol. 52, no. 12, pp. 3314–3323, December 2004.

[6] P. M. T. Broersen, "Finite sample criteria for autoregressive order selection," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3550–3558, 2000.

[7] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, 1st ed., ser. Information Science and Statistics. Springer, 2005.

[8] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[9] ——, *Information and Complexity in Statistical Modeling*, 1st ed., ser. Information Science and Statistics. Springer, 2007.

[10] V. Balasubramanian, "MDL, Bayesian inference, and the geometry of the space of probability distributions," in *Advances in Minimum Description Length: Theory and Applications*, I. J. M. P. D. Grünwald and M. A. Pitt, Eds. MIT Press, 2005, pp. 81–99.

[11] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1712–1717, July 2001.

[12] ——, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40–47, January 1996.

[13] C. D. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," *IMS Lecture Notes–Monograph Series, Time Series and Related Topics*, vol. 52, pp. 48–59, 2006.

[14] A. I. McLeod and Y. Zhang, "Partial autocorrelation parameterization for subset autoregression," *Journal of Time Series Analysis*, vol. 27, no. 4, pp. 599–612, 2006.

[15] B. Porat and B. Friedlander, "Computation of the exact information matrix of Gaussian time series with stationary random components," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 118–130, 1986.

[16] P. M. Broersen, *Automatic Autocorrelation and Spectral Analysis*, 1st ed. Springer, June 2006.

[17] S. M. Kay, "Recursive maximum likelihood estimation of autoregressive processes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, pp. 56–65, 1983.

[18] M. C. Jones, "Randomly choosing parameters from the stationarity and invertibility regions of autoregressive-moving average models," *Applied Statistics*, vol. 36, no. 2, pp. 134–138, 1987.

[19] J. Nocedal and S. Wright, *Numerical Optimization*, P. Glynn and S. M. Robinson, Eds. Springer, 2000.

[20] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[21] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[22] J. Rissanen and P. E. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes," *The Annals of Statistics*, vol. 7, no. 2, pp. 297–315, 1979.

[23] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[24] O. Barndorff-Nielsen and G. Schou, "On the parametrization of autoregressive models by partial autocorrelations," *Journal of Multivariate Analysis*, vol. 3, pp. 408–419, 1973.

[25] E. Makalic and D. F. Schmidt, "Fast computation of the Kullback-Leibler divergence and exact Fisher information for the first-order moving average model," *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 391–393, 2009.

[26] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, September 1946.

[27] J. Rissanen, T. Roos, and P. Myllymäki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, no. 4, pp. 839–849, 2010.

[28] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," in *Proc. 1st Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere International Center for Signal Processing, 2008, (Invited Paper).

[29] L. J. Fitzgibbon, "On sampling stationary autoregressive model parameters uniformly in r^2 value," *Statistics & Probability Letters*, vol. 76, no. 4, pp. 349–352, 2006.

[30] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 2, pp. 211–216, 2002.

[31] H. J. Newton, *Timeslab: A Time Series Analysis Laboratory*. Brooks/Cole, 1988.

[32] P. D. Tuan, "Cramer-Rao bounds for AR parameter and reflection coefficient estimators," *IEEE Transactions Acoustics, Speech and Signal Processing*, vol. 37, no. 5, pp. 769–772, 1989.