Contents lists available at ScienceDirect
Statistics and Probability Letters
journal homepage: www.elsevier.com/locate/stapro

Minimum message length analysis of multiple short time series

Daniel F. Schmidt *, Enes Makalic
Centre for Epidemiology and Biostatistics, The University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 7 September 2015
Accepted 16 September 2015
Available online xxxx

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

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.

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 Cruddas et al. (1989) 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 Cox and Solomon (1988).

* Corresponding author. Tel.: +61 3 8344 0767.
E-mail address: dschmidt@unimelb.edu.au (D.F. Schmidt).

http://dx.doi.org/10.1016/j.spl.2015.09.021

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 (Wallace, 2005) 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 (Wallace, 2005; Wallace and Boulton, 1968; Wallace and Freeman, 1987) 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 (Wallace and Boulton, 1975; Wallace and Freeman, 1987; Schmidt, 2011). The most commonly used approximation in the minimum message length literature is the Wallace–Freeman, or MML87, codelength approximation (Wallace and Freeman, 1987). 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 normalised second moment of an optimal quantising lattice in k-dimensions (Conway and Sloane, 1998). Following (Wallace, 2005, 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 minimises (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 (Dowe and Wallace, 1997) and factor analysis (Wallace and Freeman, 1992; Wallace, 2005, 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),