ISSN 1440-771X
0 7326 1099 0

MONASH
UNIVERSITY
AUSTRALIA

DEPARTMENT OF ECONOMETRICS AND BUSINESS STATISTICS

Model Selection Criteria for Segmented Time Series from a Bayesian Approach to Information Compression

Brian Hanlon and Catherine Forbes

Working Paper 8/2002

Model Selection Criteria for Segmented Time Series from a Bayesian Approach to Information Compression

Brian Hanlon[*1] *& Catherine Forbes*[2]

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

[1]Brian.Hanlon@dsto.defence.gov.au
[2]Catherine.Forbes@BusEco.monash.edu.au

Abstract

The principle that the simplest model capable of describing observed phenomena should also correspond to the best description has long been a guiding rule of inference. In this paper a Bayesian approach to formally implementing this principle is employed to develop model selection criteria for detecting structural change in financial and economic time series. Model selection criteria which allow for multiple structural breaks and which seek the optimal model order and parameter choices within regimes are derived. Comparative simulations against other popular information based model selection criteria are performed. Application of the derived criteria are also made to example financial and economic time series.

Keywords: Complexity theory, segmentation, break points, change points, model selection, model choice.

JEL Classification: C11.

1. Introduction

It is well known that models with constant coefficients provide poor descriptions of time series which contain structural change [Maddala & Kim 2000]. While structural changes in time series may be gradual, sudden shifts must also be encompassed as possible exogenous events. A failure to accommodate such structural breaks could result in misleading conclusions, particularly with respect to unit root tests [Perron 1989]. In this paper we develop model selection criteria which test for multiple abrupt break points in time series. The number and location of the break points are, *a priori*, unknown. The order of models and the model parameters within each regime are also, *a priori*, unknown.

The developed model selection criteria are based on a particular information principle with strong foundations in the fields of complexity theory and computability. Like other information based criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), a balance is sought between the information cost of describing the model against that of describing the data relative to the model. The principle employed originates with Wallace and Freeman [1987] and is known as Minimum Message Length (MML) compact coding. This principle shares many similarities with Rissanen's Minimum Description Length (MDL) approach [Rissanen 1989] which is finding growing application within statistics [Hansen & Yu 2001]. Importantly, however, MML and MDL differ in key areas, chief amongst which is MML's embrace of an explicitly Bayesian formulation. This provides MML with some

[*] Permanent address: Aeronautical & Maritime Research Laboratory, Air Operations Division, 506 Lorimer Street, Fishermans Bend 3207, Victoria, Australia

immediate advantages for model selection: focus can be directed to the selection of particular models rather than model classes and model selection criteria remain invariant under general parameter transformations.

In this paper MML based model selection criteria are derived for segmented time series with regions described by (i) simple Gaussian models and (ii) Gaussian auto-regressive models. The criteria allow for the simultaneous selection of the change points, the order of each auto-regressive model, and all other model parameters.

As MML is relatively unknown amongst econometricians Sections 2 & 3 will outline the foundations of MML and its computational framework. In Section 4, model selection criteria for determining multiple structural breaks in a simple (uncorrelated) time series are developed. A small simulation study is presented in Section 5 to demonstrate the performance of the derived criteria. Comparisons with other information based model selection criteria suggest that the MML criteria developed here provide a good balance between over and under-fitting on the number of segment boundaries. In addition, the Gaussian multiple segmentation approach is demonstrated on Gold Bullion and Brent crude oil price series. These ideas are fully extended in the case of the autoregressive (AR) multiple segmentation problem, presented in Section 6. A demonstration of the approach on the application of finding a structural break in the quarterly US GDP is given. The paper concludes with a discussion in Section 7.

2. Complexity, Information Theory and the Minimum Message Length

Although largely developed within computer science, MML is suitable for application to econometrics as it rests on the assumption that probabilistic models provide a sufficient class from which descriptions of complicated data generating mechanisms can be drawn. Probabilistic models provide the basis for a pragmatic way of computing the complexity of a given time series and its generating mechanism. Complexity relates to how difficult it is to describe the time series. Basing time series on probabilistic models, and thus a random variable, introduces the assumption that the time series are intrinsically difficult to describe - each data point must be described individually as it cannot be deterministically related to other data points in the series. However, any probabilistic model introduces additional structure which may moderate this indeterminism. It is for this reason that the measure of complexity for a time series must incorporate how difficult it is to describe the data *and* the model.

The pragmatic approach to calculating complexity is based on information theory. Utilising the framework of information theory, descriptions of time series can be couched in the language of minimising the length of messages; this is the origin of the Minimum Message Length principle. MML thus requires the construction of a message, the length of which determines an objective function called a message length. The idea is to envisage the sending of a ‘message’ relaying the precise values of an observed data series. A model is considered useful when sending both the parameter estimates and the residuals from the fitted model results in a message that is shorter than any other model from some class. The task of model selection then relates to finding the model which results in the shortest message. As the message length combines both a measure of complexity in the model with that in the data, the MML approach provides a method for simultaneously evaluating the trade off between the two. Adjustments are made to ensure that the methodology is robust against minor variations in how the message is constructed. In regular problems, MML point estimators have good properties, such as asymptotic consistency, and are

closely related to both the Maximum Likelihood Estimators (MLE) and Bayesian posterior mode estimators [Wallace & Freeman 1987].

As with AIC and BIC, MML information criteria or "Message Lengths" (MessLen) have the general form:

$$MessLen = -\log(likelihood) + \text{penalty term}. \quad (1)$$

However, in contrast to AIC and BIC, the MML penalty term, for regular problems, is not a simple function of the number of parameters, d, and the sample size, n, but rather has the structure:

$$\text{penalty term} = -\log(prior) + \frac{1}{2}\log(\det F(\theta)) + g(d). \quad (2)$$

Here $\log(prior)$ refers to the logarithm of the prior density over the parameters, θ, $\det(F(\theta))$ is the determinant of the Fisher information matrix, and $g(d)$ is a model specific function of the number of parameters. Importantly, as the penalty term is a function of the unknown parameters, θ, the MML information criteria are not typically evaluated at the MLE.

What can be gained by including a complicated penalty function of the form in Equation 2? Consider first the inclusion of the prior density function. This allows additional contextual information which may otherwise be abstracted out of the model selection criterion to be explicitly incorporated. The term involving the Fisher information matrix captures the fact that, for some models, certain regions of the parameter space are structurally more informative than others. That is, regardless of the observed data, some parameter value regions are more likely to maximise the likelihood function. Conversely, as the minimum message length is sought, regions that are relatively more informative are penalised within the MML approach. A tension is thus established between minimising the message length and providing sufficient support to describe the data. Without the Fisher information matrix term this structural information on candidate models would be ignored.

When the problem contains non-regular components, such as when estimating unknown multiple structural breaks, the penalty term can become even more complicated. However, the additional complication in the criteria serve to balance the complexity in the models in much the same way.

MML information criteria are explicitly Bayesian model selection criteria but are in contrast to the standard Bayesian approach which involves the use of the marginal probabilities of models, conditional on the observed data [Chibb 1995]. Standard Bayesian model selection will typically first chose the "best" model, and subsequently minimise a loss function to obtain a parameter estimate for the chosen model. Such procedures typically ignore the uncertainty in model choice when presenting the subsequent parameter estimates. A popular approach involves calculating posterior model probabilities which are then used to construct an "averaged model" that marginalises over the uncertainty in the model choice [Geweke 1999]. The model averaged approach is particularly relevant for prediction purposes. For the problem of determining the number and occasions of structural breaks in a time series, however, it is worthwhile to explore the use of choosing a single model, or potentially best set of models, favouring differing structural change patterns. MML offers the possibility of simultaneously evaluating the uncertainty in determining multiple

structural breaks with that of estimating parameter values for the resulting individual models.

Earlier attempts to derive MML based model selection criteria for application to the uni-variate segmentation problem, where each segment is defined by a simple Gaussian model [Oliver, Baxter & Wallace 1998] [Baxter & Oliver 1996], have failed to take full advantage of the discrete nature of the segmentation problem. This paper seeks to overcome this limitation and to derive robust MML based model selection criteria which: (1) Make full use of the information inherent in the data sets; (2) Are consistent with known constraints on devising descriptive messages; and (3) Are amenable to robust application under generalised computational search algorithms.

3. The Minimum Message Length Computational Framework

Wallace and Freeman [1987] demonstrate the principles of MML inductive inference. Under certain regularity conditions, the MML objective function is shown to take the generic form:

$$MessLen = L(\theta) - \log h(\theta) + \frac{1}{2}\log(\det F(\theta)) + \frac{d}{2}(1 + \log \kappa_d). \qquad (3)$$

Here $h(\theta)$ is the density of the prior distribution for θ and κ_d is the d dimensional quantising lattice constant [Conway & Sloane 1982] required for optimal encoding of the d dimensional θ. Values of κ_d for dimensions one to eight can be found in Appendix 1. Parameter values which minimise the message length are inferred to be those which are the most suitable to describe the data from that class and order of model. We briefly outline the main derivation of Equation 3.

Consider a parametric statistical model for a set of data *D,* $f(D \mid \theta)$. The negative log-likelihood is then given by $L(\theta) = -\log f(D \mid \theta)$, which reflects the size of the message required to describe the data. The base of the logarithm corresponds to the number of symbols in the coding alphabet used to construct the message. For convenience natural logarithms are often used, so that messages are measured in 'nits' rather than the usual base 2 'bits'. Strictly, the negative log-likelihood corresponds to the data message length only when the probability density is multiplied with the uncertainty in the data value, resulting in a probability function. The probability density is assumed to be sufficiently well behaved to allow for this approximation. The length of the message is thus made dependent on the accuracy of the data. However, as this is usually constant for all the data points this data uncertainty can be scaled out of the MessLen.

For the vector of parameters, θ, $h(\theta)$ represents the Bayesian prior probability distribution over the parameters. From Shannon and Weaver's work on information entropy [Shannon & Weaver 1959], it follows that the message length for the parameters will be given by $-\log Vh(\theta)$. V corresponds to the parameter space uncertainty volume. As with the data, the Bayesian prior is assumed sufficiently well behaved such that $Vh(\theta)$ approximates the local probability. Unlike with the data, the parameter space uncertainty volume need not be constant for all parameter values and thus cannot be scaled out.

The introduction of parameter uncertainty partitions the parameter space into discrete values centred on the regions of uncertainty. However, this partitioning is not unique. To avoid variations in the message length from variations in how the parameter

space is partitioned consideration is given to the expected message length. The message length for the data, averaged over the parameter space uncertainty volume is given by:

$$MessLen(D \mid \theta) = \frac{1}{V} \int_V L(\theta + \bar{\theta}) dv \tag{4}$$

where $\bar{\theta}$ represents the vector of parameter deviations from the optimal vector of parameter values. Assuming that the distribution $f(D \mid \theta)$ is regular, the message length for the data can be Taylor expanded. Further assuming that the estimate of the parameter values is unbiased, so that the expectation of $\bar{\theta}$ is zero, the Taylor expanded message length of the parameters and data becomes:

$$MessLen(\theta \,\&\, D) = -\log Vh(\theta) + L(\theta) + \frac{1}{2V} \int_V \bar{\theta}^T \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \bar{\theta} dV \tag{5}$$

To minimise Equation 5, the optimal parameter values and uncertainty volume need to be determined. The tricky part concerns the minimisation of the parameter uncertainty, represented by the volume V. Specifically, it turns out that the optimal value for V is a function of the integrand in Equation 5. As a consequence, the optimal value for V is a function of the data leading to a problem of circular dependence so that the transmitted message cannot be decoded. A solution to this problem is to integrate out the data dependence and approximate the integrand of Equation 5 with the expected Fisher information:

$$\frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \approx F(\theta) = \int f(D \mid \theta) \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} dy \tag{6}$$

where y represents the vector of data values. Evaluating the integral and optimising Equation 5 with respect to the parameter space uncertainty volume yields a message length for both the data and the parameter values as given in Equation 3, which can be seen to have a penalty function of the form in Equation 2.

4. MML Model Selection Criteria for Simple Segmented Gaussian Models

4.1 Notation

To apply MML to the segmentation problem the same notation utilised in [Oliver, Baxter & Wallace 1998] is employed. Data consists of pairs (x_i, y_i) with the x_i evenly spaced over a region of size R. The region is presumed known and can be cut into $C+1$ pieces by C segment boundaries, or change points. The change points are denoted by $\{v_1, v_2, ..., v_C\}$. It is assumed throughout that errors are Gaussian. The observations x_i in segment i are assumed to have constant mean and variance, c_i and σ_i^2, respectively.

Since the x_i are evenly spaced, and particularly as they form a finite set, it will prove convenient to map them onto the set of integers **Z** corresponding to the sequence of the data points.

θ' corresponds to the set of parameters having continuous support. θ' will thus not include the change point parameters.

4.2 The case with one change point

It will be shown that the MML information criterion for the simplest case of determining the location of a single change point is given by

$$MessLen(\theta \,\&\, data) \approx -\log(h(\theta')) - \log(\frac{1}{R}) + \frac{1}{2}\log\left(\det F(\theta')\right) - \log\left(s+1\right) + L_0 + L_1 + \frac{s\left(\frac{s}{2}+1\right)}{8\left(s+1\right)}\left(\frac{\sigma_1^2 - \sigma_0^2 + D^2}{\sigma_0^2} + \frac{\sigma_0^2 - \sigma_1^2 + D^2}{\sigma_1^2}\right) + 2 + 2\log\kappa_4 \tag{7}$$

where L_0 and L_1 are the negative log likelihood's for segments 0 and 1, $D = c_0 - c_1$ and $s \in \{0,2,4,6,...\}$ is a discrete parameter reflecting the size of the uncertainty in the change point position.

To derive Equation 7, we note that, as with other model parameters, change points in an MML description must be accompanied by some non-zero uncertainty as to their location. Let ε denote the difference in the location of the change point defined at the centre of the region of uncertainty, $\hat{v}$, and the maximum likelihood value, v:

$$\varepsilon = \hat{v} - v \tag{8}$$

Denoting by n_0 and n_1 the number of data items in segments 0 and 1, as defined by v, the number of data items, n, can be represented with respect to $\hat{v}$ as:

$$n \;=\; \begin{array}{l}\text{Number of data}\\ \text{items in Segment 0}\\ \text{as defined by } \hat{v}\end{array} \;+\; \begin{array}{l}\text{Number of data}\\ \text{items in Segment 1}\\ \text{as defined by } \hat{v}\end{array}$$

which can be expressed as:

$$n = \text{int}\left(v + \frac{n\varepsilon}{R}\right) + \left(n_0 + n_1 - \text{int}\left(v + \frac{n\varepsilon}{R}\right)\right) \tag{9}$$

where the operator $\text{int}()$ returns the value of the operand truncated to its integer value. By definition $\text{int}(v) = n_0$. Note that by mapping the x_i onto **Z** and measuring R as the number of data points in its range the data density $\frac{n}{R}$ will be unity.

Given that a range of v will lead to the *same* segmentation of the data, the issue arises as to what values v should actually take. This can be ascertained by noting that the x_i are model parameters already endowed with a parameter accuracy. With

the x_i mapped onto **Z,** as noted above, a parameter accuracy of one unit is implicit with the midpoint of the region centred on x_i. Neither the sender nor receiver can have greater information than this. Change points are thus naturally defined as occurring mid-way between the x_i parameters and also endowed with an underlying parameter accuracy of one unit. Variations of the change point position within this range lead to the same segmentation. Should minimisation of the message length demand that the location of the change point parameter be described less accurately, the region of uncertainty for a change point must necessarily be in multiples of the underlying parameter accuracy. Put succinctly, the change point positions have only a discrete set of points for support, not the entire set of real numbers. Writing $Acc_v = s + 1$, and given the underlying discrete support of the change points, it follows that any additional uncertainty in the change point parameter value can only be expressed by one of $\frac{s}{2} \in \{0, 1, 2, ...\}$, as shown in Figure 1:

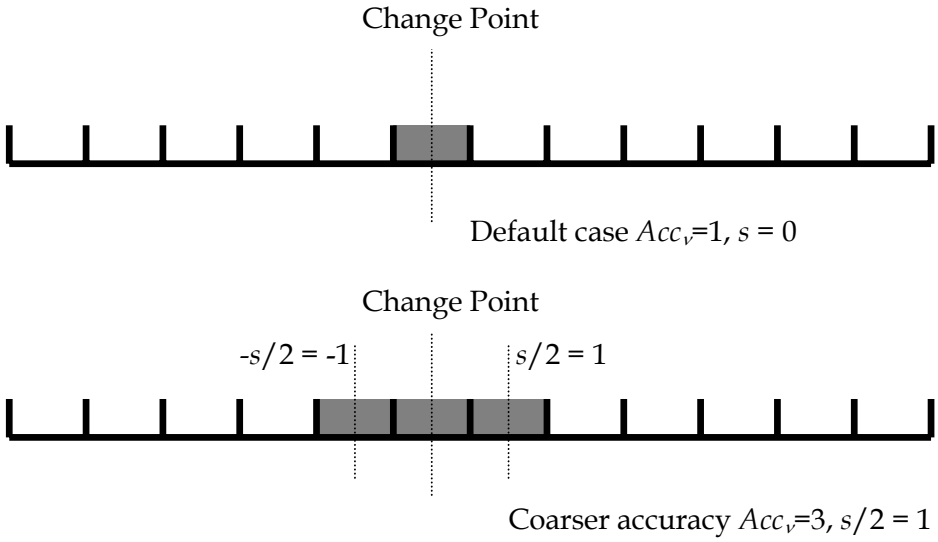

Figure 1. Two examples showing the relation between the discrete parameter s and the Acc_v: (1) a case where no additional parameter accuracy is needed and (2) when the cut point parameter is to be transmitted with less accuracy, in this case with the next lowest accuracy.

It follows that $s \in \{0,2,4,6.....\}$. As required, s is a multiple factor on the underlying parameter accuracy, assumed known a priori by the receiver of the message. Indeed, as the x_i are treated as *data* their accuracy of measurement must be presumed known. Note that minimisation of the message length may not require that the change point location be described with courser accuracy. In this case $Acc_v = 1$ and s will not appear in the message length description.

This formulation differs from [Oliver, Baxter & Wallace 1998] and [Baxter & Oliver 1996] in that s is precluded from taking arbitrary values and the change points occur only mid-way between the x_i. It follows that now $\varepsilon \in \{0,\pm 1,\pm 2,...\}$ so that Equation 9 becomes:

$$\begin{aligned} n &= \text{int}(v + \varepsilon) + (n_0 + n_1 - \text{int}(v + \varepsilon)) \\ &= (n_0 + \varepsilon) + (n_1 - \varepsilon) \end{aligned} \qquad (10)$$

The message length description of the data with respect to the quantised estimate of the change point will be given by:

$$L(\theta) = (n_0 + \varepsilon)\log\left(\sqrt{2\pi}\sigma_0\right) + \sum_{i=1}^{n_0+\varepsilon}\frac{(y_i - c_0)^2}{2\sigma_0^2} + (n_1 - \varepsilon)\log\left(\sqrt{2\pi}\sigma_1\right) + \sum_{i=n_0+\varepsilon+1}^{n}\frac{(y_i - c_1)^2}{2\sigma_1^2} \tag{11}$$

where θ represents the model parameters. Writing $h(\theta)$ for the prior on the model parameters, the message length on the parameters and the data can be generically represented as [Wallace & Freeman 1987]:

$$MessLen(\theta \,\&\, data) \approx -\log(V_5 h(\theta)) + \frac{1}{V_5}\int_{V_5}\left(L(\theta) + \overline{\theta}\frac{\partial L(\theta)}{\partial\theta} + \frac{1}{2}\overline{\theta}^T\frac{\partial^2 L(\theta)}{\partial\theta^T\partial\theta}\overline{\theta}\right)dV_5 \tag{12}$$

where $\overline{\theta}$ corresponds to a parameter perturbation and V_5 is the five dimensional volume element made up from the four continuous and one change point parameter uncertainties. Since the Fisher information is not defined for change point like parameters the volume element pertaining to the change point uncertainty is presumed orthogonal to the parameter uncertainties which make up the remaining volume element. The message length description of Equation 12 can thus be written as:

$$\begin{aligned} MessLen(\theta \,\&\, data) \approx\; & -\log((s+1)h(v)) - \log(V_4 h(\theta')) \\ & + \frac{1}{(s+1)V_4}\sum_{\varepsilon=-\frac{s}{2}}^{\frac{s}{2}}\int_{V_4}\left(L(\theta) + \overline{\theta'}\frac{\partial L(\theta)}{\partial\theta'} + \frac{1}{2}\overline{\theta'}^T\frac{\partial^2 L(\theta)}{\partial\theta'^T\partial\theta'}\overline{\theta'}\right)dV_4 \end{aligned} \tag{13}$$

where θ' refers to the parameters other than the change point parameter. Note that the expansion now is defined with respect to the differentiable parameters only. As with the standard MML formulation the quantised estimates of the continuous parameters are presumed to be unbiased so that the linear term in the Taylor expansion of Equation 13 will vanish.

Determining the expected message length with respect to the change point parameter thus requires attention to be focussed on both the $L(\theta)$ and $\frac{\partial^2 L(\theta)}{\partial\theta^T\partial\theta}$ terms. Only the first term was modified by the presence of change points in [Oliver, Baxter & Wallace 1998] and [Baxter & Oliver 1996]. For both terms the logarithmic contributions add nothing new to the message length in expectation. Non-trivial contributions to the message length description arise through the sums in Equation 11. The expectations with respect to the change point parameter of these sums can be determined easily and are found to be:

$$E_{cut}\left(\sum_{i=1}^{n_0+\varepsilon}\frac{(y_i-c_0)^2}{2\sigma_0^2}\right)=\sum_{i=1}^{n_0}\frac{(y_i-c_0)^2}{2\sigma_0^2}+\frac{1}{(s+1)}\sum_{\varepsilon=1}^{\frac{s}{2}}\left(\sum_{i=n_0+1}^{n_0+\varepsilon}\frac{(y_i-c_0)^2}{2\sigma_0^2}-\sum_{i=n_0+1-\varepsilon}^{n_0}\frac{(y_i-c_0)^2}{2\sigma_0^2}\right) \tag{14}$$

and

$$E_{cut}\left(\sum_{i=n_0+\varepsilon+1}^{n}\frac{(y_i-c_1)^2}{2\sigma_1^2}\right)=\sum_{i=n_0+1}^{n}\frac{(y_i-c_1)^2}{2\sigma_1^2}+\frac{1}{(s+1)}\sum_{\varepsilon=1}^{\frac{s}{2}}\left(\sum_{i=n_0-\varepsilon+1}^{n_0}\frac{(y_i-c_1)^2}{2\sigma_1^2}-\sum_{i=n_0+1}^{n_0+\varepsilon}\frac{(y_i-c_1)^2}{2\sigma_1^2}\right) \tag{15}$$

The expectation over $L(\theta)$ with respect to the change point parameter can thus be expressed as:

$$E_{cut}(L(\theta))=L_0+L_1+$$
$$+\frac{1}{2(s+1)}\sum_{\varepsilon=1}^{\frac{s}{2}}\left(\sum_{i=n_0+1}^{n_0+\varepsilon}\left\{\frac{(y_i-c_0)^2}{\sigma_0^2}-\frac{(y_i-c_1)^2}{\sigma_1^2}\right\}+\sum_{i=n_0+1-\varepsilon}^{n_0}\left\{\frac{(y_i-c_1)^2}{\sigma_1^2}-\frac{(y_i-c_0)^2}{\sigma_0^2}\right\}\right) \tag{16}$$

where L_0 and L_1 are the negative log likelihood's for segments 0 and 1 respectively. Note that the effect of the segmentation uncertainty is restricted to the data items actually within the region of uncertainty and not, as in [Oliver, Baxter & Wallace 1998] and [Oliver & Forbes 1997], dependent on all the data items.

The $\frac{\partial^2 L(\theta)}{\partial\theta^T\partial\theta}$ term can be calculated similarly. To make the message decodeable, this term is approximated by its expectation over the data, i.e. the Fisher information. Expressing the Fisher information matrix as:

$$F(\theta')=\begin{bmatrix} I_{c_0c_0} & I_{c_0c_1} & I_{c_0\sigma_0} & I_{c_0\sigma_0} \\ I_{c_1c_0} & I_{c_1c_1} & I_{c_1\sigma_0} & I_{c_1\sigma_1} \\ I_{\sigma_0c_0} & I_{\sigma_0c_1} & I_{\sigma_0\sigma_0} & I_{\sigma_0\sigma_1} \\ I_{\sigma_1c_0} & I_{\sigma_1c_1} & I_{\sigma_1\sigma_0} & I_{\sigma_1\sigma_1} \end{bmatrix} \tag{17}$$

the non-vanishing contributions are found to be given by:

$$I_{c_0c_0}=\frac{n_0}{\sigma_0^2};\quad I_{c_1c_1}=\frac{n_1}{\sigma_1^2};\quad I_{c_0\sigma_0}=I_{\sigma_0c_0}=\frac{s\left(\frac{s}{2}+1\right)(c_1-c_0)}{2(s+1)\sigma_0^3};\quad I_{c_1\sigma_1}=I_{\sigma_1c_1}=\frac{s\left(\frac{s}{2}+1\right)(c_0-c_1)}{2(s+1)\sigma_1^3}$$
$$I_{\sigma_0\sigma_0}=\frac{2n_0}{\sigma_0^2}+\frac{3s\left(\frac{s}{2}+1\right)\left(\sigma_1^2-\sigma_0^2+D^2\right)}{4(s+1)\sigma_0^4};\quad I_{\sigma_1\sigma_1}=\frac{2n_1}{\sigma_1^2}+\frac{3s\left(\frac{s}{2}+1\right)\left(\sigma_0^2-\sigma_1^2+D^2\right)}{4(s+1)\sigma_1^4}. \tag{18}$$

It is seen that the imprecision in the change point leads to some mixing of parameters from the two regions in the Fisher information matrix, as would be expected.

Given the vanishing contributions, the determinant of the Fisher matrix can be expressed as:

$$\det\left(F(\theta')\right) = \prod_{j\in\{1,2\}} \left(I_{c_j c_j} I_{\sigma_j \sigma_j} - I^2_{\sigma_j c_j} \right) \tag{19}$$

Allowing for the case that the data is not segmented, there are R possible positions for the change point. The Wallace-Freeman form for the message length [Wallace & Freeman 1987] for both the parameters and data is then given by (where the change point positions are presumed uniformly distributed):

$$\begin{aligned} MessLen(\theta \,\&\, data) \approx & -\log(h(\theta')) - \log(\frac{1}{R}) + \frac{1}{2}\log\left(\det F(\theta')\right) - \log\left(s+1\right) + L_0 + L_1 \\ & + \frac{1}{2(s+1)} \sum_{\varepsilon=1}^{\frac{s}{2}} \left(\sum_{i=n_0+1}^{n_0+\varepsilon} \left\{ \frac{(y_i - c_0)^2}{\sigma_0^2} - \frac{(y_i - c_1)^2}{\sigma_1^2} \right\} + \sum_{i=n_0+1-\varepsilon}^{n_0} \left\{ \frac{(y_i - c_1)^2}{\sigma_1^2} - \frac{(y_i - c_0)^2}{\sigma_0^2} \right\} \right) + 2 + 2\log\kappa_4 \end{aligned} \tag{20}$$

where κ_4 is the four dimensional quantising lattice constant.

Isolating terms which are a function of s, it is clear that the optimal value of s will be a function of the data. In order that the message be decodeable, parameter uncertainties must be expressed in a data independent way. The simplest way to achieve this is to extend the approximation used by Wallace and Freeman [1987] and approximate the data dependent term:

$$\frac{1}{2(s+1)} \sum_{\varepsilon=1}^{\frac{s}{2}} \left(\sum_{i=n_0+1}^{n_0+\varepsilon} \left\{ \frac{(y_i - c_0)^2}{\sigma_0^2} - \frac{(y_i - c_1)^2}{\sigma_1^2} \right\} + \sum_{i=n_0+1-\varepsilon}^{n_0} \left\{ \frac{(y_i - c_1)^2}{\sigma_1^2} - \frac{(y_i - c_0)^2}{\sigma_0^2} \right\} \right) \tag{21}$$

by its expectation. The Minimum Message Length description is then given by Equation 7.

It is to be noted that expressing the minimal value for the change point precision in a data independent way has not be emphasised previously [Oliver, Baxter & Wallace 1998] and [Oliver & Forbes 1997]. Note also that the additional terms to the Minimum Message Length arising from a coarser segmentation accuracy vanish if $s = 0$. These contributions also vanish if the means and variances of the two regions are equal.

4.3 The MML approach to multiple change points

The only additional complication introduced by considering multiple change points concerns accurately accounting for segments which are bordered by change points at both ends. Proceeding as before it is found that this aspect of the problem contributes new structure to the Fisher matrix in a regular way. Generalising from Equation 21, an expression for the message length with an arbitrary number of change points can be directly written down:

$$MessLen(\theta \,\&\, data) \approx -\sum_{j=0}^{C} \log(h(c_j, \sigma_j)) - C\log\left(\frac{1}{R}\right) + \frac{1}{2}\log \prod_{j \in Segments} \left(I_{c_j c_j} I_{\sigma_j \sigma_j} - I_{\sigma_j c_i}^2\right)$$
$$- \sum_{j=0}^{C-1} \log(s_j + 1) + \sum_{j=0}^{C} L_j + \sum_{j=0}^{C-1} MixingTerms_{j,j+1} - \log C! + \frac{d}{2} + \frac{d}{2}\log \kappa_d \qquad (22)$$

where

$$MixingTerms_{j,j+1} = \frac{s_j\left(\frac{s_j}{2}+1\right)}{8(s_j+1)}\left(\frac{\sigma_{j+1}^2 - \sigma_j^2 + D_{j,j+1}^2}{\sigma_j^2} + \frac{\sigma_j^2 - \sigma_{j+11}^2 + D_{j,j+1}^2}{\sigma_{j+1}^2}\right) \qquad (23)$$

and the non-vanishing contributions to the Fisher matrix for each segment are given by:

$$I_{c_j c_j} = \frac{n_j}{\sigma_j^2}$$

$$I_{c_j \sigma_j} = I_{\sigma_j c_j} = \sum_{i \in \, bordering \; segments} \frac{s_i\left(\frac{s_i}{2}+1\right)(c_i - c_j)}{2(s_i+1)\sigma_i^3}$$

$$I_{\sigma_j \sigma_j} = \frac{2n_j}{\sigma_j^2} + \sum_{i \in \, bordering \; segments} \frac{3s_i\left(\frac{s_i}{2}+1\right)\left(\sigma_i^2 - \sigma_j^2 + D_{ij}{}^2\right)}{4(s_i+1)\sigma_j^4} \qquad (24)$$

As before, d is the dimension of the precision volume for the continuous parameters.

5. Applications to Simple Gaussian Segmentation

5.1 Small sample comparative study

The MML approach to segmented data series advocated here is clearly more complex than other model selection criteria. It is thus necessary to test if the additional effort involved in deriving such a criterion is balanced by its utility. To investigate this, data series were generated which consisted of two segments. Each segment was assigned the same variance, 0.5, but different means. The difference in the means between the segments was systematically reduced to investigate the ability of the MML criterion to select the correct number of segments.

To undertake this experiment a simulation program was constructed which employs the down-hill simplex algorithm of Nelder and Mead [Press, Teukolsky, Vetterling & Flannery 1993] to find the parameter values and uncertainties which minimise the message length. As with the simulations performed by Baxter and Oliver [1996], single change point solutions which yield local minima in the message length were determined first. Change point locations across the entire data range were tested and those yielding the deepest local minima retained. Multiple change point solutions based on the retained single change point locations, yielding up to three segments,

were then investigated. Solutions comprising no, one or two change points were then compared and the optimal choice determined. For each separation of the two "true" segments, fifty data samples were recorded. Each data series consisted of fifty data points with the true segmentation at data point eighteen.

This experiment was applied to the derived MML criterion as well as the Akaike Information Criterion (AIC) and the Minimum Description Length (MDL) criterion. Comparison is made with the AIC and MDL criteria defined by [Dom 1995] and [Liang, Jaszczak & Coleman 1992]:

$$AIC = -\log f(y \mid \theta) + number\ of\ parameters$$
$$MDL = -\log f(y \mid \theta) + \frac{continuous\ parameters}{2}\log n + \log\binom{n}{C} \qquad (25)$$

Simple priors for the continuous parameters were chosen based on the population estimates:

$$h(\sigma) = \frac{1}{\sigma_{population}}, \qquad \frac{\sigma_{population}}{2} < \sigma < \frac{3\sigma_{population}}{2}$$

$$h(\mu) = \frac{1}{2\sigma_{population}}, \qquad \mu_{population} - \sigma_{population} < \mu < \mu_{population} + \sigma_{population} \qquad (26)$$

The observed percentage chance of identifying one change point (that is one segment boundary) for different differences in the segment means is shown in Figure 2:

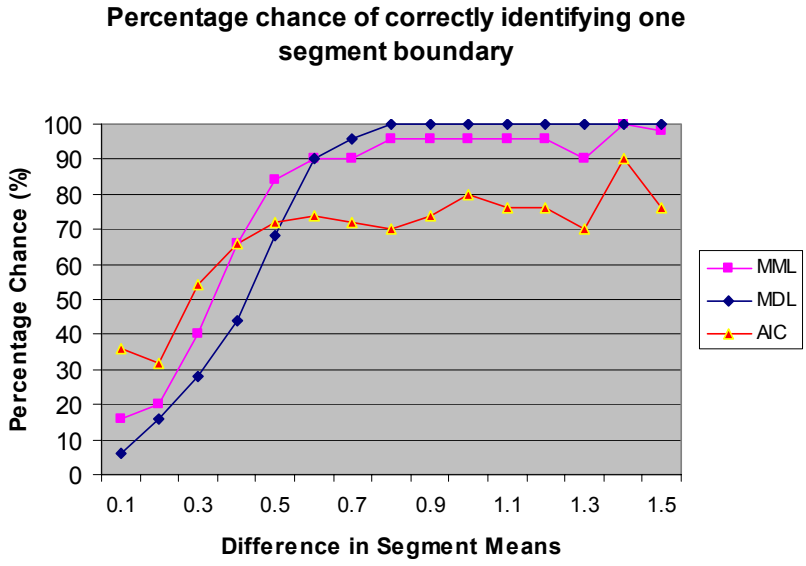

Figure 2. Percentage chance of identifying one segment boundary for different differences in the segment means for the MML, AIC and MDL criteria.

As can be seen from Figure 2, the MML criterion performs well across the range of mean distances, demonstrating greater sensitivity to the existence of change points

than MDL. This is highlighted in Table 1 which logs the propensity for the different model selection criteria to overfit and underfit by giving the observed percentage chance of finding three or one segments rather than two. Table 1 also records the quality of the segments chosen through use of a simple distance measure between the chosen and true segment:

$$\text{Change Point Position Accuracy} = |\ \text{Found Segment Position} - \text{True Segment Position}\ |$$

The Change Point Position Accuracy results are averaged over the cases when a single change point was found for each model selection criterion.

Delta Mean	Average Change Point Position Accuracy			% Chance of overfitting			% Chance of underfitting		
	MML	MDL	AIC	MML	MDL	AIC	MML	MDL	AIC
1.5	0.204	0.22	0.211	2	0	24	0	0	0
1.4	0.28	0.28	0.267	0	0	10	0	0	0
1.3	0.267	0.26	0.286	10	0	30	0	0	0
1.2	0.313	0.34	0.289	4	0	24	0	0	0
1.1	0.583	0.56	0.684	4	0	24	0	0	0
1.0	0.5	0.56	0.475	4	0	20	0	0	0
0.9	0.792	0.76	0.784	4	0	26	0	0	0
0.8	0.938	1.1	1.0	4	0	30	0	0	0
0.7	1.51	1.71	1.39	10	0	28	0	4	0
0.6	1.91	2.36	2.35	8	0	26	2	10	0
0.5	3.74	4.35	3.64	6	0	26	10	32	2
0.4	4.76	4.59	4.12	12	0	20	22	56	14
0.3	7.1	6.71	8.52	10	0	26	50	72	20
0.2	7.8	5.75	7.19	8	0	24	72	84	44
0.1	15.13	15.67	12.22	8	0	18	76	94	46

Table 1. The average Change Point Position Accuracy and percentage chance of over-fitting and underfitting for the different model selection criteria for different differences in the segment means.

The percentage chance of over-fitting demonstrates the well known over-fitting properties of AIC. Conversely, MDL shows no propensity to overfit, all the incorrect choices for the number of segments coming from under-fitting. MML presents less tendency to underfitting than the MDL criterion for small mean differences but retains a tendency to overfit at larger differences, emphasising its sensitivity to the existence of segment boundaries. The MML criterion thus demonstrates a better balance between over and under-fitting. Importantly, this implies that the MML criterion will be a preferred choice if seeking to detect segment boundaries in noisy data.

Interestingly, the average Change Point Position Accuracy does not vary greatly between criteria, implying that if the correct number of change points can be found they will be located in the data sequence equally well.

5.2 Behaviour of the MML criterion for increasing sample size

Since the MML criterion is not optimised at the MLE and the change points are discrete parameters, it is difficult to formally demonstrate that the MessLen has the appropriate asymptotic properties [McQuarrie & Tsai 1998]. However, the trend behaviour of the MML criterion in detecting the correct number of segments for increasing sample size can be demonstrated empirically, as shown in Figure 3. In these simulations equal numbers of additional data points were added to each “true” region on either side of the change point. As shown in Figure 3, the percentage

chance of identifying one change point (or segment boundary) shows the anticipated asymptotic behaviour.

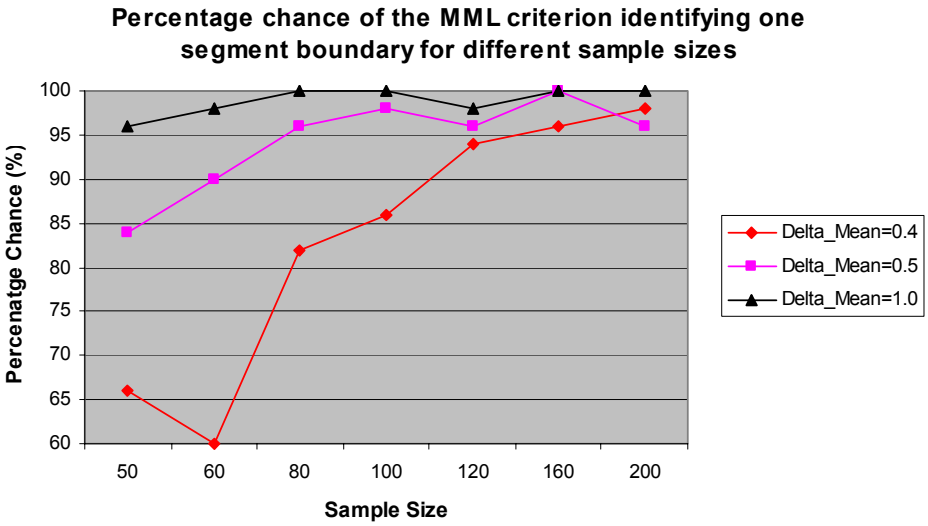

Figure 3. Percentage chance of identifying one segment boundary using MML for different sample sizes and differences in the segment means.

5.3 Application to economic time series

Utilising the prior of Equation 26, the MML criterion was applied to two sets of economic time series. The time series correspond to the Brent crude oil price and Gold bullion price in the interval 8th September to 16th of November 1999. This time period was chosen as it appeared to offer the possibility of multiple segments in both time series. Pricing data were examined as they can demonstrate regions of relatively stable behaviour so that simple Gaussian distributions may be employed as reasonable models for the time series. This is to be contrasted with data such as national GDP which generally shows a monotonic rise. The segmentations found are shown in Figures 4 & 5:

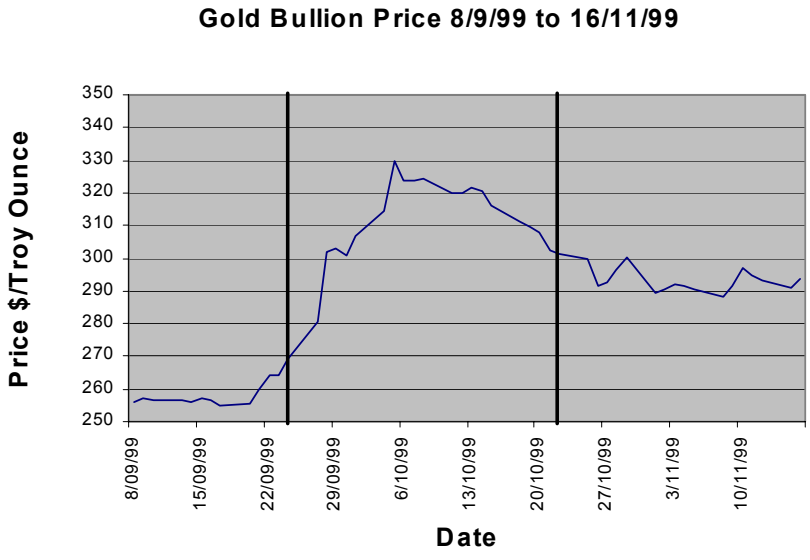

Figure 4. Gold bullion price between 8th September to 16th November 1999, showing the derived MML segment boundaries.

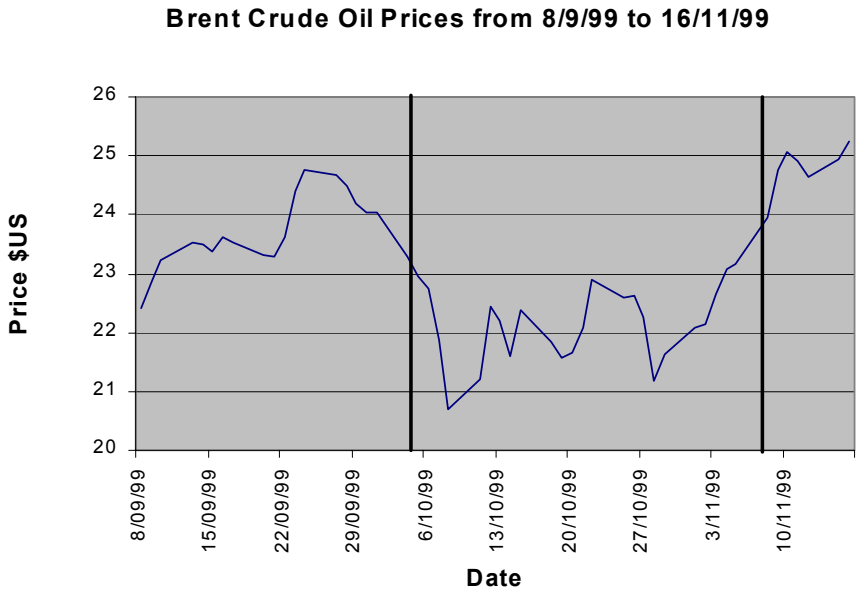

Figure 5. Brent crude oil price between 8th September to 16th November 1999, showing the derived MML segment boundaries.

The same algorithm used in section 5 was employed. The existence of up to three segments was tested. In both figures 3 and 4 the segment boundaries chosen by the MML criterion appear to accord with behavioural changes in the data sets.

6. MML Model Selection Criteria for Segmented Gaussian Auto-regressive Models

6.1 Background

In this section extension is made to the segmentation of time series where different segments are described by auto-regressive, AR(p), models of, perhaps different, order p. The AR(p) models are represented in the generic form:

$$y_t = \sum_{i=1}^{p} \rho_i y_{t-i} + c + \varepsilon_t \qquad (27)$$

Application of auto-regressive models is perhaps one of the most popular for time series data. By extending consideration to the segmentation of auto-regressive models the utility of applying MML to the analysis of structural change in a broader range of time series can be tested. The greatest difficulty in undertaking an analysis of the application of MML to the segmentation of auto-regressive time series comes from the computational difficulties which arise in dealing with the expectations which must be calculated. Such complications arise when seeking an MML description of auto-regressive models beyond first order and in the MML description of segmented auto-regressive models of any order. It is to be recalled that the need to calculate expectations was directly linked with the ability to determine the optimal uncertainty volume in a way which made the MML description viable.

Rather than rely on expectations, an alternative approach is to add an additional preamble to the message which explicitly transmits the size of the uncertainty volume. The MML description is thus made more complex through this approach but such a prescription seems unavoidable if MML is to be applied to more realistic time

series models [Wallace & Freeman 1987]. Consistent with the other parameters, including a preamble for the uncertainty volume requires that an uncertainty in the uncertainty volume also be considered. An infinite regress of uncertainties is avoided by demonstrating that this uncertainty in uncertainty is in fact a function of the original uncertainty volume, so that the message can be made intelligible. It was stated in [Wallace & Freeman 1987] that the uncertainty in the *logarithm* of the uncertainty volume in the case of a single parameter model has size $\sqrt{6}$. In this section this result will be generalised to models consisting of *d* continuous parameters. Extension will then be made to segmented time series.

As it is amenable to direct calculation, the standard MML description of first order auto-regressive models will be given in the Appendix 2.

6.2 MML description of auto-regressive models

Adding the third component describing the size of the parameter space uncertainty, Equation 4 generalises to:

$$\begin{aligned} MessLen(\theta \,\&\, D) = & -\log\left(V_{\log V} h_2(\log V)\right) \\ & + \frac{1}{V_{\log V}} \int_{V_{\log V}} -\log V h(\theta) + L(\theta) + \frac{1}{2V} \int_V \bar{\theta}^T \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \bar{\theta} dV dV_{\log V} \end{aligned} \quad (28)$$

where, to be consistent with the treatment in [Wallace & Freeman 1987], $V_{\log V}$ is the uncertainty in the logarithm of the parameter uncertainty and $h_2(\log V)$ is the prior probability of realising the logarithm of the uncertainty volume V. Consideration is given to logarithms as this preserves the general form of the message length.

Similarly to the standard MML treatment, consideration is given to the message length averaged over the uncertainty in the logarithm of the parameter space uncertainty. To demonstrate that an infinite regress can be avoided the optimal value for $V_{\log V}$ needs to be determined. To evaluate the integral an appropriate parameter space transformation is introduced which renders the volume V into a *d* dimensional sphere of volume U [Oliver & Baxter 1995]. However, unlike in the standard MML approach, the approximation of integrating out the data dependence is not pursued. After introducing this transformation and evaluating the integral over V, Equation 28 becomes:

$$\begin{aligned} & MessLen(\theta \,\&\, D) \\ & = -\log V_{\log V} h_2(\log V) + L(\theta) + \frac{1}{V_{\log V}} \int_{V_{\log V}} -\log U g(\phi) + \frac{d}{2} \kappa_d U^{\frac{2}{d}} dV_{\log V} \end{aligned} \quad (29)$$

where $g(\phi)$ is the transformed prior over parameter values. To evaluate the remaining integral the integrand must first be expanded around the optimal value for $\log U$. Similarly, again, with the standard MML approach, assume that the expectation of $\log U$ over the volume $V_{\log V}$ is unbiased and that the second moment behaves as from a uniform distribution. Expanding the integrand in Equation 29 and utilising this assumption yields the message length:

$$MessLen(\theta \,\&\, D) = -\log V_{\log V} h_2(\log V) + L(\theta) - \log U g(\phi) + \frac{d}{2}\kappa_d U^{\frac{2}{d}} + \frac{V_{\log V}^2}{12d}\kappa_d U^{\frac{2}{d}} \tag{30}$$

Optimising Equation 30 with respect to $V_{\log V}$ yields:

$$V_{\log V} = \sqrt{\frac{6d}{\kappa_d U^{\frac{2}{d}}}} \tag{31}$$

Substituting the optimal value for $V_{\log V}$, Equation 31, and transforming yields:

$$MessLen(\theta \,\&\, D) = -\log V_V g_V(V) + L(\theta) - \log V h(\theta) + \frac{d}{2}\kappa_d \left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta}\right)^{\frac{1}{d}} V^{\frac{2}{d}} + \frac{1}{2} \tag{32}$$

where $g_V(V)$ is the prior in the parameter space uncertainty. Equation 32 can be reduced further by noting that $V_V = V V_{\log V}$. Equation 32 thus becomes:

$$\begin{aligned} MessLen(\theta \,\&\, D) = &-\log \left(\frac{6d}{\kappa_d \left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta}\right)^{\frac{1}{d}} V^{\frac{2}{d}}} \right)^{\frac{1}{2}} V - \log g_V(V) \\ &+ L(\theta) - \log V h(\theta) + \frac{d}{2}\kappa_d \left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta}\right)^{\frac{1}{d}} V^{\frac{2}{d}} + \frac{1}{2} \end{aligned} \tag{33}$$

Differentiating Equation 33 with respect to the parameter space uncertainty results in the following expression for the optimal parameter space uncertainty volume:

$$\left(\frac{1-2d}{d}\right)\frac{1}{V} - \frac{1}{g_V(V)}\frac{\partial g_V(V)}{\partial V} + \kappa_d \left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta}\right)^{\frac{1}{d}} V^{\frac{2}{d}-1} = 0 \tag{34}$$

It follows from Equation 34 that the optimal parameter space uncertainty volume will be an explicit function of the parameter space uncertainty prior. This dependence can be overcome by restricting the prior to be uniform. This does not, in the first instance, seem unreasonable as we are attempting to introduce uncertainty into uncertainty, the form of which we will generally be ignorant. There are some interesting parallels in taking such a stance with the notion of hierarchical priors, where a second stage subjective prior is placed on a prior capturing structural knowledge. It is often the case that the second stage prior is chosen to be a suitable non-informative prior [Berger 1985]. Nevertheless, in the MML context, introducing this assumption runs contrary to the general philosophy of MML.

With this constraint on the form of the parameter space uncertainty prior, the optimal parameter space uncertainty becomes:

$$V = \left(\frac{2d-1}{\kappa_d d} \right)^{\frac{d}{2}} \frac{1}{\left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \right)^{\frac{1}{2}}} \tag{35}$$

or noting that $U = \left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \right)^{\frac{1}{2}} V$, the optimal value for U becomes:

$$U = \left(\frac{2d-1}{\kappa_d d} \right)^{\frac{d}{2}} \tag{36}$$

which is equivalent to the standard MML result when $d = 1$. The optimal value for the uncertainty in the logarithm of the parameter space uncertainty is then given by:

$$V_{\log V} = \sqrt{\frac{6d^2}{2d-1}} \tag{37}$$

which recovers the result in Wallace and Freeman (1987) when $d = 1$. Making the relevant substitutions into Equation 33, the Message Length becomes:

$$MessLen(\theta \,\&\, D) =$$

$$-\frac{1}{2}\log\left(\frac{6d^2}{2d-1} \right) - \log\left(\frac{g_V(V)}{\left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \right)^{\frac{1}{2}}} \right) + L(\theta) - d\log\left(\frac{2d-1}{\kappa_d d} \right) - \log\left(\frac{h(\theta)}{\left(\det \frac{\partial^2 L(\theta)}{\partial \theta^T \partial \theta} \right)^{\frac{1}{2}}} \right) + d \tag{38}$$

6.3 MML description of segmented auto-regressive models

Consideration is now given to applying this generalised formulation of the extended MML preamble to higher order segmented AR(*p*) models. Starting with the case of a model with a single change point dividing AR(*l*) and AR(*m*) descriptions of each segment, the negative log-likelihood is expressed as:

$$\begin{aligned} L(\theta) = (n_l + \varepsilon)\log\left(\sqrt{2\pi}\sigma_l\right) + \sum_{i=l}^{n_l+\varepsilon} \frac{\left(y_i - \sum_{j=1}^{l} \rho_j y_{i-j} - c_l \right)^2}{2\sigma_l^2} \\ + (n_m - \varepsilon)\log\left(\sqrt{2\pi}\sigma_m\right) + \sum_{i=n_l+\varepsilon+1+m}^{n} \frac{\left(y_i - \sum_{j=1}^{m} \beta_j y_{i-j} - c_m \right)^2}{2\sigma_m^2} \end{aligned} \tag{39}$$

where the total number of data points, n, can be expressed as before as $n = (n_l + \varepsilon) + (n_m - \varepsilon)$, with n_l and n_m the number of data points in each segment, with respect to the change point position which minimises the MML description, and ε is the discrete uncertainty in the position of the change point. Generalising the formulation of Section 4 and introducing the uncertainty in the change point position and the uncertainty in the uncertainty of the change point position, Equation 28 generalises to:

$$MessLen(\theta \,\&\, D) = -\log(s_s + 1) g_s(s_0) - \log V_{\log V} h_V(\log V) + \frac{1}{V_{\log V}(s_s + 1)} \times$$
$$\sum_{\frac{s}{2}=\frac{s_0}{2}-\frac{s_s}{2}}^{\frac{s_0}{2}+\frac{s_s}{2}} \int_{V \log_V} -\log V h(\theta) - \log(s+1) h_s(\nu) + \frac{1}{(s+1)V} \sum_{\varepsilon=-\frac{s}{2}}^{\frac{s}{2}} \int_V L(\theta) + \frac{1}{2} \bar{\theta}^T \frac{\partial^2 L(\theta)}{\partial \theta^2} \bar{\theta} d\nu V dV_{\log V} \tag{40}$$

where s_s is the uncertainty in the change point position uncertainty, s_0 is the change point uncertainty which minimises the message length and $g_s(s_0)$ is the prior probability of realising a given uncertainty in the change point position uncertainty. Note that $s_0 \geq s_s$. Expressing the negative log-likelihood averaged over change point uncertainties as:

$$\mathbf{L}(\theta, s_0, s_s) = \frac{1}{(s_s + 1)} \sum_{\frac{s}{2}=\frac{s_0}{2}-\frac{s_s}{2}}^{\frac{s_0}{2}+\frac{s_s}{2}} \frac{1}{(s+1)} \sum_{\varepsilon=-\frac{s}{2}}^{\frac{s}{2}} L(\theta) \tag{41}$$

Equation 40 can be expressed as:

$$MessLen(\theta \,\&\, D) = -\log(s_s + 1) g_s(s_0) - \log V_{\log V} h_V(\log V)$$
$$- \frac{1}{(s_s + 1)} \sum_{\frac{s}{2}=\frac{s_0}{2}-\frac{s_s}{2}}^{\frac{s_0}{2}+\frac{s_s}{2}} \log(s+1) h_s(\nu) + \frac{1}{V_{\log V}} \int_{V \log V} -\log V h_\theta(\theta) + \frac{1}{V} \int_V \mathbf{L}(\theta, s_0, s_s) dV dV_{\log V} \tag{42}$$

With the addition of the terms involving s_0 and s_s, Equation 42 has the same form (once the generalised negative log-likelihood, Equation 41, is Taylor expanded) as Equation 28. The analysis leading to the optimal parameter uncertainty volume and optimal uncertainty in the uncertainty volume is thus the same with the substitution of the negative log-likelihood with Equation 41.

One element to the message length which has been overlooked thus far is the contribution from the conditional data values upon which the models of the data depend. For segmented auto-regressive models, conditional data points will be required for each segment, the number being equal to the order of the auto-regressive model. Such conditional data points thus appear as parameters but are clearly different to the model parameters which are being sought through MML inference. The differences arise from the observation that: (1) The Accuracy of Measurement (AOM) of the data is presumed known so that there is no need to optimise against the uncertainty volume of these “parameters”; and (2) The

quantisation of the data is presumed known so that there is no need to average over the AOM of the conditional data points.

To transmit these conditional data points the segmented auto-regressive models cannot be used and we are forced to adopt a more crude approach. The simplest way to achieve this is to make use of the data range and transmit each conditional data point with a message of length $-\log\frac{AOM}{Data\_Range}$, where AOM is the accuracy to which the data is given, presumed known. Noting that the AOM is scaled out of the message length the message length for the segmented AR(*l*)-AR(*m*) model takes the form:

$$MessLen(\theta \,\&\, D) = -\log(s_s+1) g_s(s_0) - \frac{1}{(s_s+1)} \sum_{\frac{s}{2}=\frac{s_0}{2}-\frac{s_s}{2}}^{\frac{s_0}{2}+\frac{s_s}{2}} \log(s+1) h_s(\nu)$$

$$-\frac{1}{2}\log\left(\frac{6d^2}{2d-1}\right) - \log\left(\frac{g_V(V)}{\left(\det\frac{\partial^2 \mathbf{L}(\theta, s_0, s_s)}{\partial\theta^T \partial\theta}\right)^{\frac{1}{2}}}\right) - \log\left(\frac{h(\theta)}{\left(\det\frac{\partial^2 \mathbf{L}(\theta, s_0, s_s)}{\partial\theta^T \partial\theta}\right)^{\frac{1}{2}}}\right) \qquad (43)$$

$$+\mathbf{L}(\theta, s_0, s_s) - d\log\left(\frac{2d-1}{\kappa_d d}\right) + d - (m+l)\log\frac{1}{Data\_Range}$$

where now d is the dimensionality of the continuous parameters from both segments. The message length in the non-segmented case, Equation 38, needs to be similarly modified to account for the conditional data points. It is straightforward to determine the non-vanishing terms to the determinant in Equation 43 from the definition of $\mathbf{L}(\theta, s_0, s_s)$ given in Equation 41. Similarly, by generalising Equations 39 and 42 to an arbitrary number of segments, and by adding the requisite number of prior probabilities, Equation 43 can be simply extended to the description of auto-regressive models with an arbitrary number of change points.

6.4 Application to US GDP data

6.4.1 Background

In testing the application of this MML approach to segmented auto-regressive models, it is clearly possible to undertake a general search over model classes which allow for auto-regressive models of arbitrary order in each segment. To simplify the analysis, consideration is here restricted to searching for a single change point in time series where each segment is described by an AR(1) model, so connecting analysis with the current literature on unit roots. We consider the US quarterly GDP between 1960:III and 1985:II, so encompassing the 1973 oil price shock. As in [Nelson & Plosser 1982] consideration is given to the data transformed to natural logs.

Following the approach in Section 5, application is made of an optimisation model utilising the down-hill simplex method [Press et al] to search for parameter estimates which minimise the MML selection criterion. Several searches were performed with the boundary conditions on the prior probabilities implemented with increasing

severity. In this way different local minima were sometimes revealed and the deepest one retained.

6.4.2 Choice of priors

The choice of priors will be guided by allowing for the possibility of both a unit root and a structural break. Following Marriot and Newbold [1998 & 2000] we employ a beta prior for the autoregressive parameter in each segment, so introducing increasing probability mass about the unit root. The prior(s) take the form:

$$h(\rho_i) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)2^{\alpha+\beta-1}}(1+\rho_i)^{\alpha-1}(1-\rho_i)^{\beta-1}, \quad |\rho_i| < 1 \tag{44}$$

Marriot and Newbold [1998] obtained attractive results when the hyper-parameters take the values $\alpha = 5$ and $\beta = 0.5$. These values will be employed here.

Continuing with the prior belief that the different segments are described by a unit root, we assume that each segment is difference stationary and introduce a simple Gaussian prior for the mean values. This is based on population estimates of the mean and variance of the first differenced data:

$$h(c_i) \sim N(c_{\Delta y}, \sigma_{\Delta y}) \tag{45}$$

To allow for unbounded values of the standard deviation but with increasingly small probability as deviation is made from the sample value we employ a Gamma distribution:

$$h(\sigma_i) = \frac{1}{\gamma^{\delta}\Gamma(\delta)}\sigma_i^{\delta-1}e^{\frac{\sigma_i}{\gamma}}, \quad \sigma_i > 0 \tag{46}$$

with the hyper-parameters set to $\delta = 1$ and $\gamma = \sigma_{\Delta y}$ as in [Oliver & Forbes 1997].

The prior belief in the uncertainty in the change point location uncertainty is taken to be the same as the usual prior over change point positions, i.e. $\frac{1}{R}$, where R is the number of data points. The prior for the uncertainty in the uncertainty volume for the continuous parameters is more subtle. Recalling that this prior must be uniform in the uncertainty volume, the simplest choice appears to be to simply utilise Equation 35 and set:

$$g_V(V) = \frac{\left(\det \frac{\partial^2 \mathbf{L}}{\partial\theta^T\partial\theta}\right)^{\frac{1}{2}}}{\left(\frac{2d-1}{\kappa_d d}\right)^{\frac{d}{2}}} \equiv \frac{1}{V_0} \qquad \frac{V_0}{2} < V < \frac{3V_0}{2} \tag{47}$$

This choice is somewhat incestuous but has the benefit of: (1) Providing an objective prior for a parameter which is of no direct interest to the statistician interested in the

problem of inferring the model parameters; and (2) Simplifying the message length description to a form originally envisaged in [Wallace & Freeman 1987].

The second point can be illustrated by considering the case when there are no change points. Utilising Equation 47, the MML description, Equation 38 with the conditional data point contribution, becomes:

$$MessLen(\theta \,\&\, D) = -\frac{1}{2}\log\left(\frac{6d^2}{2d-1}\right) - \frac{d}{2}\log\left(\frac{2d-1}{\kappa_d d}\right) + L(\theta) - \log\left(\frac{h(\theta)}{\left(\det\frac{\partial^2 L(\theta)}{\partial\theta^T\partial\theta}\right)^{\frac{1}{2}}}\right) + d - m\log\frac{1}{Data\_Range} \tag{48}$$

for an auto-regressive model of order *m*, which is similar to the standard MML description with the addition of some terms which are a function of d. Indeed, it follows that when $d = 1$ the contribution from the extra preamble adds a small contribution "of order one or so" [Wallace & Freeman 1987].

6.4.3 Results

The minimum message lengths found when the change-point was located at different positions of the time series are shown in Figure 6:

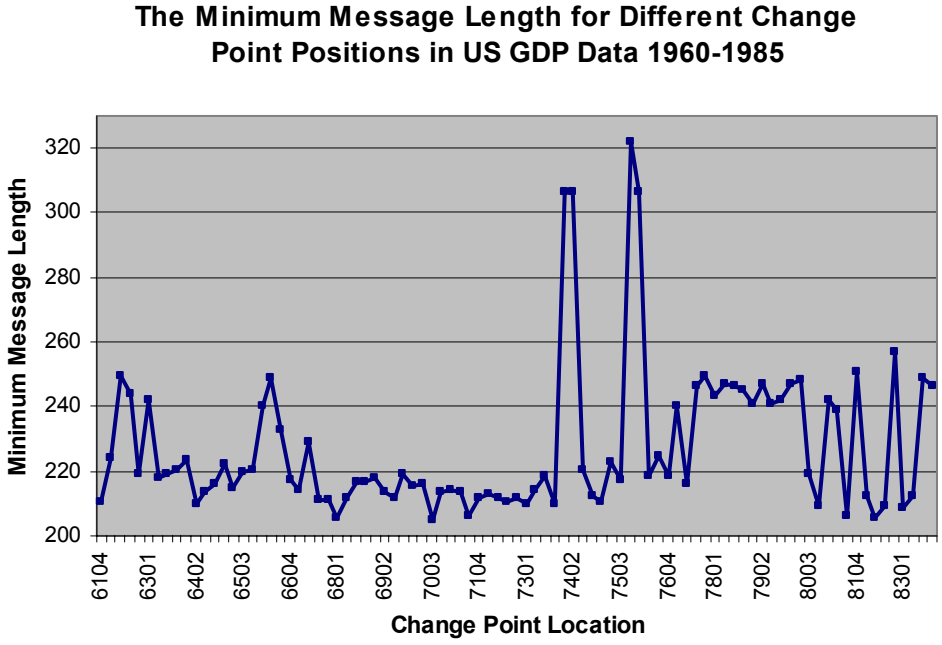

Figure 6: The found minimum message length for US GDP data for different locations of the change point.

The minimum solution was found with a change point located at the transition from 1970:III to 1970:IV with a message of length 204.8 nits, narrowly beating the solution

when no change points are present, which yielded a message of length 209.0 nits. The optimal expected fit is shown in Figure 7:

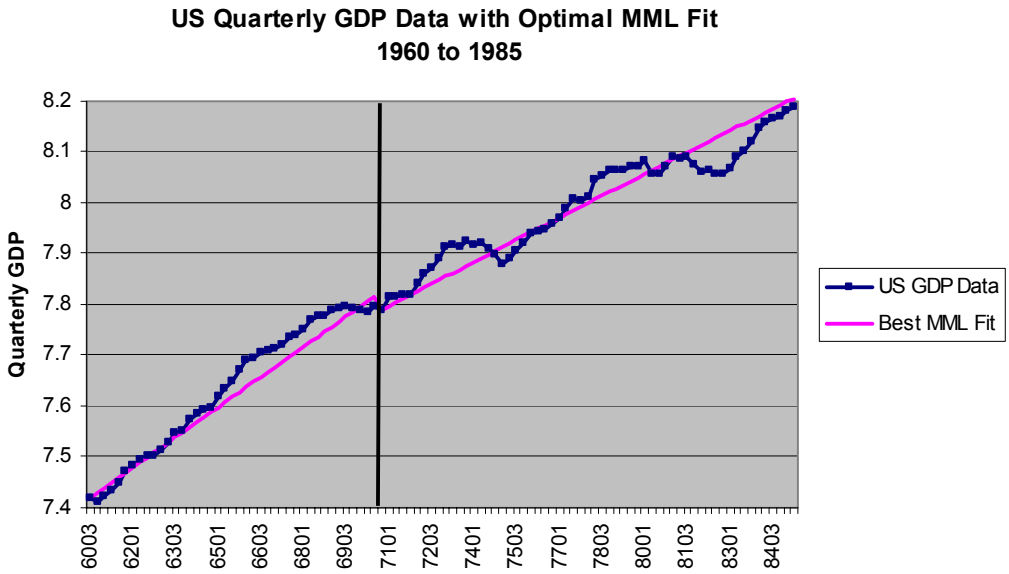

Figure 7: US quarterly GDP with optimal MML fit, showing a change point at the 1970:III – 1970:IV transition.

The parameter values which minimise the MML criterion are:

$$\begin{aligned} \rho_0 &= 1.000000 & \rho_1 &= 0.999233 \\ c_0 &= 0.009902 & c_1 &= 0.013314 \\ \sigma_0 &= 0.010138 & \sigma_1 &= 0.012191 \end{aligned} \tag{49}$$

It is worth noting that several minima were found which out-performed the no change point solution.

As suggested by Perron [1989] we detect a change in the slope of the trend but slightly earlier than the supposed 1973:I date for the oil shock. We also find a small change in the intercept.

7. Conclusion

The approach to segmentation advocated in this paper has sought to formulate a representation of the segmentation problem in a way which is consistent with the general prescriptions of MML and the information actually present in the, discrete, data. In doing so, the effects of segmentation appear as structure in the main body of the message length and in the Fisher information matrix. This additional structure in the Fisher information matrix has not been accommodated in previous MML segmentation criteria.

There are several important advantages to formulating segmentation criteria in this way which are explicitly highlighted through MML: (1) The change point positions are treated discretely; (2) Only change point positions which yield different segmentations are utilised; (3) Only discrete numbers of change points are affected by change point uncertainties; (4) The pathological case of allowing a change point to

lie on a data item is excluded; (5) The change point uncertainty can never be less than that originally given in the data; (6) Should the original change point uncertainty optimise the estimate, any additional structure in the estimator arising from change point uncertainties vanishes; and (7) Multiple change points are not independent.

In the case of simple Gaussian models it has been demonstrated that the MML criteria developed here demonstrate good properties of discrimination when seeking segment boundaries. In the case of Gaussian auto-regressive models it has been demonstrated that MML criteria can be developed and applied to the segmentation problem by generalising an alternative prescription to overcome computational complexities, yielding interesting results.

As with any optimisation problem, however, the greatest difficulty comes from discriminating global from local minima in a computational search. As suggested by Maddala & Kim [2000] it may be more profitable to concentrate the search for change points in a local area of the time series rather than broadly as implemented here. This would certainly allow computational efforts to be focussed more directly on the issue of finding global minima.

8. Acknowledgments

It is a pleasure to thank Prof. Paul Kofman, School of Finance and Economics, University of Technology, Sydney, for providing the economic data upon which some of the simulations were based. Special thanks also to Prof. Essie Maasoumi for comments on the manuscript.

Appendix 1: Quantising Lattice Constants

The quantising lattice constant is known exactly for the first eight dimensions. These values are given in Table 2:

Dimension d	Lattice Constant κ_d
1	0.083333
2	0.080188
3	0.078543
4	0.076603
5	0.075625
6	0.074244
7	0.073116
8	0.071682

Table 2. The Quantising Lattice Constants for Dimensions One to Eight.

For arbitrary dimensions the quantising lattice constant is unknown but is known to obey the bounds shown in Equation 50 [Conway & Sloane 1982]:

$$\frac{\Gamma\left(\frac{d}{2}+1\right)^{\frac{2}{d}}\Gamma\left(1+\frac{2}{d}\right)}{d\pi} > \kappa_d > \frac{\Gamma\left(\frac{d}{2}+1\right)^{\frac{2}{d}}}{(d+2)\pi} \tag{50}$$

Both bounds approach $\frac{1}{2\pi e}$ as d increases.

Appendix 2: The Standard MML Description of AR(1) Models

Consideration is directed to the standard MML description of non-segmented AR(1) models. The negative log likelihood function in the case that $p = 1$ is given by:

$$L(\theta) = n\log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \rho y_{t-1} - c)^2 \tag{51}$$

To determine the relevant expectations, it is convenient to first recursively solve for y_t. This allows the data to be expressed as functions of the model parameters and Gaussian errors only. The resulting expressions are found to be:

$$y_t = \begin{cases} \rho^t y_0 + c\left(\dfrac{1-\rho^t}{1-\rho}\right) + \displaystyle\sum_{i=0}^{t-1}\varepsilon_{t-i}\rho^i & \rho \neq 1 \\ & t > 0 \\ y_0 + tc + \displaystyle\sum_{i=0}^{t-1}\varepsilon_{t-i} & \rho = 1 \end{cases} \tag{52}$$

which are conditional on y_0. Expressing the Fisher information matrix as

$$F(\theta) = \begin{bmatrix} I_{\rho\rho} & I_{\rho\sigma} & I_{\rho c} \\ I_{\sigma\rho} & I_{\sigma\sigma} & I_{\sigma c} \\ I_{c\rho} & I_{c\sigma} & I_{cc} \end{bmatrix} \tag{53}$$

where $I_{\theta\theta'} = E\left(\partial^2 L(\theta) \big/ \partial\theta\partial\theta'\right)$, the non-vanishing terms can be determined by utilising Equation 52 and noting that terms linear in the Gaussian errors will vanish. It is found that the non-vanishing terms are given by:

$$I_{\rho c} = \begin{cases} \dfrac{1}{\sigma^2}\left[(\rho y_0 - \dfrac{c\rho}{1-\rho})\left(\dfrac{1-\rho^{n-1}}{1-\rho}\right) + \dfrac{c(n-1)}{1-\rho} + y_0\right] & \rho \neq 1 \\ \dfrac{ny_0}{\sigma^2} + \dfrac{c}{\sigma^2}\dfrac{n(n-1)}{2} & \rho = 1 \end{cases} \tag{54}$$

and

$$I_{\rho\rho} = \begin{cases} (\frac{y_0}{\sigma})^2\left(\frac{1-\rho^{2n}}{1-\rho^2}\right)+\frac{n}{1-\rho^2}-\frac{1-\rho^{2n}}{(1-\rho^2)^2} \\ +(\frac{c}{\sigma})^2\left(\frac{n}{(1-\rho)^2}+\frac{1-\rho^{2n}}{(1-\rho)^2(1-\rho^2)}-\frac{2(1-\rho^n)}{(1-\rho)^3}\right) \\ +(\frac{2y_0c}{\sigma^2})\left(\frac{1-\rho^n}{(1-\rho_j)^2}-\frac{1-\rho^{2n}}{(1-\rho)(1-\rho^2)}\right) & \rho \neq 1 \\ \\ \frac{ny_0}{\sigma^2}+\left(\frac{2y_0c}{\sigma^2}+1\right)\frac{n(n-1)}{2}+\left(\frac{c}{\sigma}\right)^2\frac{n(n-1)(2n-1)}{6} & \rho = 1 \end{cases} \tag{55}$$

with $I_{\sigma\sigma} = \frac{2n}{\sigma^2}$ and $I_{cc} = \frac{n}{\sigma^2}$. Note that Equation 55 reduces to the results found in [Phillips 1991] when $c = 0$.

Noting that there are three parameters, the message length for the AR(1) model can now be written down directly as:

$$\begin{aligned} MessLen(\theta \,\&\, data) = &-\log h(\theta) + n\log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(y_t - \rho y_{t-1} - c\right)^2 \\ &+\frac{1}{2}\log\left(\det F(\theta)\right) + \frac{3}{2} + \frac{3}{2}\log\kappa_3 - \log\frac{1}{Data\_Range} \end{aligned} \tag{56}$$

where θ represents the vector of model parameters and $h(\theta)$ is the prior for those parameters.

References

Baxter, R. A. & Oliver, J. J., The Kindest Cut: Minimum Message Length Segmentation*, Lecture Notes in Artificial Intelligence 1160, Algorithmic Learning Theory*, edited by S. Arikawa and A. Sharma, Springer-Verlag Berlin, pg 83, 1996.

Berger, J. O., *Statistical Decision Theory and Bayesian Analysis,* Springer-Verlag, New York, 1985.

Chib, S., Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association,* Volume 90, page 1313, 1995.

Conway, J. H. and Sloane, N. J. A, Voronoi Regions on Lattices, Second Moments of Polytopes and Quantization, *IEEE Trans. Inf. Thy, IT-28*, Page 211, 1982.

Dom, B., *MDL Estimation with Small Sample Sizes Including an Application to the Problem of Segmenting Binary Strings using Bernoulli Models*, Technical Report RJ9997 (89085) IBM Research Division, Almaden Research Center, San-Jose California, 1995.

Geweke, J., Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication, *Econometric Reviews*, Volume 18, page 1, 1999.

Hansen, M. & Yu, B., Model Selection and the Principle of Minimum Description Length, *Journal of the American Statistical Association,* Volume 96, page 746, 2001.

Liang, Z., Jaszczak, R. J. and Coleman, R. E., Parameter Estimation of Finite Mixtures using EM Algorithm and Information Criteria with Applications to Medical Image Processing, *IEEE Transactions on Nuclear Science,* Volume 39, Page 1126, 1992.

Maddala, G. S. & Kim, I., *Unit Roots, Cointegration and Structural Change,* Cambridge University Press, United Kingdom, 2000.

Marriot, J. & Newbold, P., Bayesian Comparison of ARIMA and stationary ARMA Models, *International Statistical Review,* Volume 66, Page 323, 1998.

Marriot, J. & Newbold, P., The Strength of Evidence for Unit Autoregressive Roots and Structural Breaks: A Bayesian Perspective, *Journal of Econometrics,* Volume 98, Page 1, 2000.

McQuarrie, A. D. R and Tsai, C., *Regression and Time Series Model Selection*, World Scientific, Singapore, 1998.

Nelson, C. R. & Plosser, C. I., Trends and Random Walks in Macroeconomic Time Series, *Journal of Monetary Economics,* Volume 10, Page 132, 1982.

Oliver, J. J. and Baxter, R. A., *MML and Bayesianism: Similarities and Differences*, Monash University Technical Report, 1995.

Oliver,J. J., Baxter, R. A., & Wallace, C. S., Minimum Message Length Segmentation, *Research and Development in Knowledge Discovery and Data Mining (PAKKD-98)*, Springer, Berlin, Page 83, 1998.

Oliver, J. J. and Forbes, C. S., *Bayesian Approaches to Segmenting a Simple Time Series*, Monash University Working Paper 14, 1997.

Perron, P., The Great Crash, The Oil Price Shock, and the Unit Root Hypothesis, *Econometrica,* Volume 57, page 1361, 1989.

Philips, P. C. B., To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends, *Journal of Applied Econometrics,* Volume 6, Page 333, 1991.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P., *Numerical Recipes in C*, Cambridge University Press, 1993.

Rissanen, J., *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

Shannon, C. E. & Weaver, W., *The Mathematical Theory of Communication,* Urbana, University of Illinois Press, 1959.

Wallace, C. S. & Freeman, P. R., Estimation and Inference by Compact Coding, *Journal of the Royal Statistical Society (Series B)* Volume 49, page 240, 1987.