ISSN 1032-3813
ISBN 0 7326 1037 0

MONASH UNIVERSITY

AUSTRALIA

BAYESIAN APPROACHES TO SEGMENTING A SIMPLE TIME SERIES

Jonathan J. Oliver
Catherine S. Forbes

Working Paper 14/97
November 1997

DEPARTMENT OF ECONOMETRICS
AND BUSINESS STATISTICS

Bayesian Approaches to Segmenting a Simple Time Series

Jonathan J. Oliver,
Dept. of Computer Science
Monash University,
Clayton, 3168 Australia
jono@cs.monash.edu.au

Catherine S. Forbes,
Dept. Econometrics
Monash University,
Clayton, 3168 Australia
Catherine.Forbes@BusEco.monash.edu.au

November 24, 1997

Keywords: Segmentation, Minimum Message Length, MML, Bayes Factors, Evidence, Time Series

Abstract

The segmentation problem arises in many applications in data mining, A.I. and statistics. In this paper, we consider segmenting simple time series. We develop two Bayesian approaches for segmenting a time series, namely the Bayes Factor approach, and the Minimum Message Length (MML) approach. We perform simulations comparing these Bayesian approaches, and then perform a comparison with other classical approaches, namely AIC, MDL and BIC. We conclude that the MML criterion is the preferred criterion. We then apply the segmentation method to financial time series data.

1 Introduction

In this paper, we consider the problem of segmenting simple time series. We consider time series of the form:

$$y_{t+1} = y_t + \mu_j + \epsilon_t$$

where we are given N data points $(y_1 \ldots, y_N)$ and we assume there are $C+1$ segments $(j \in \{0, \ldots C\})$, and that each ϵ_t is Gaussian with mean zero and variance σ_j^2. We wish to estimate

- the number of segments, $C+1$,
- the segment boundaries, $\{v_1, \ldots, v_C\}$,
- the mean change for each segment, μ_j, and
- the variance for each segment, σ_j^2.

This model is a simplification of the TAR models proposed by Tong [24]. Figure 1 gives an example of a time series with two segments. Such models may be useful for segmenting data from (for example) (i) economic time series, (ii) electrocardiogram measurements and (iii) eye movement measurements from a sleeping person [22].

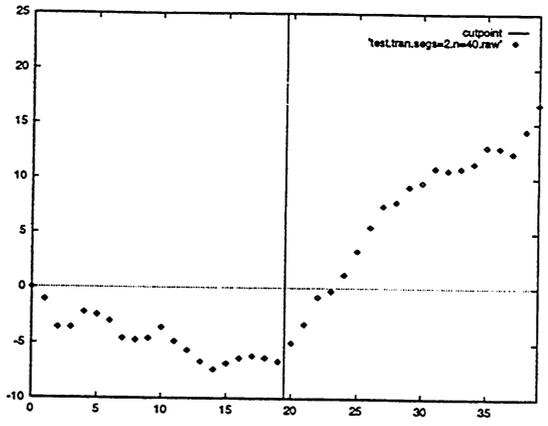

Figure 1: A time series with 2 segments

Maximum likelihood estimation is a frequently used technique for fitting the segment parameters (in our simplified case μ_j and σ_j^2). However, maximum likelihood is an inappropriate method for selecting the number of segments, since it will result in a model with homogeneous regions containing only one datum each. A variety of criteria have been used for determining the number of segments in data, including:

- Tong [24, Section 7.2.7] used AIC [1] for the segmentation of a TAR model; Liang [11] used AIC for image segmentation.

- Koop and Potter [9] use Bayes factors to compare the hypothesis of a linear model (one segment) with a single alternative of either a threshold autoregressive (TAR) model with a prespecified number of segments or a Markov Trend (or a so-called 'Hamilton') model where the number of segments is estimated using the method of Albert and Chib [2].

- Li [10] used MDL [19] (equivalent to BIC [21]) for image segmentation; Dom [7], Pfahringer [17] and Quinlan [18] refined the criterion within the context of the segmentation of binary strings.

- Baxter and Oliver [3] used MML [25, 27] (a Bayesian method for point estimation) for the segmentation of line segments with Gaussian noise.

In this paper we concentrate on the Bayesian approaches to segmentation (Bayes factors and MML). We develop the Bayesian approaches for segmentation, and highlight differences between them. We perform simulation experiments comparing the Bayesian techniques with each other, and with a variety of classical criteria. We then apply the segmentation method to financial time series data.

2 Notation

Through out this paper we describe models using the following notation:
(i) θ is the vector of continuous parameters of dimension d

$$\theta = < \mu_1, \ldots, \mu_{d/2}, \sigma_1, \ldots, \sigma_{d/2} >,$$

(ii) v is the vector of cutpoint parameters of dimension C

$$v = < v_1, \ldots, v_C >,$$

(iii) ϕ is the vector of both types of parameters of dimension $d + C$

$$\phi = < \mu_1, \ldots, \mu_{d/2}, \sigma_1, \ldots, \sigma_{d/2}, v_1, \ldots, v_C >,$$

(iv) $h(.)$ is the prior distribution function for a vector of parameters,
(v) λ, q, α and β are hyper-parameters of the prior distributions for means and standard deviations, and
(vi) j is used to index the segments; hence ($j \in \{0, 1, ,2, \ldots C\}$).

3 Bayesian Approaches for Segmentation

3.1 Prior Distributions For Parameter Values

We draw a distinction between the cutpoint parameters v and the continuous parameters θ ($\phi = \theta \cup v$). We now consider prior distributions for segmentations with C cutpoints. We assume that the prior distributions are independent:

$$h(\phi_C) = h(v)\, h(\theta_C)$$

If we let V_C be the set of possible segmentations with C cutpoints and assume that all segmentations with C cutpoints are equally likely:

$$h(v) = \frac{1}{|V_C|} \tag{1}$$

where $|V_C|$ is the number of elements in set V_C.

3.1.1 Prior Distributions For the Continuous Parameters

We consider three prior distributions over the continuous parameters μ and σ. We used Prior #1 as it made the integrations easier for the Bayes factors approach. We used two data based prior distributions (Prior #2 and Prior #3) as these prior distributions may reflect a reasonable compromise between general knowledge of an area and 'ignorance'.

For the two data based prior distributions we use the data to determine the average difference between data points and the average variability in this difference. We therefore calculate the average difference

$$\mu_{\Delta y} = \frac{\sum_{t=1}^{N-1}(y_{t+1} - y_t)}{N-1} = \frac{y_N - y_1}{N-1}$$

and the standard deviation of the differences

$$\sigma_{\Delta y} = \sqrt{\frac{\sum_{t=1}^{N-1}(y_{t+1} - y_t - \mu_{\Delta y})^2}{N-1}}$$

3.1.2 Prior #1 (A Mathematically Convenient Prior)

We assume the frequently employed mathematically convenient prior specification for each $(\mu_j,\ \sigma_j^2)$ (see Zellner [28])

$$\mu_j \mid \sigma^2 \sim N(\lambda, \sigma_j^2/q) \tag{2}$$

$$\sigma_j^2 \sim \Gamma^{-1}(\alpha, \beta) \tag{3}$$

where $\alpha,\ \beta,\ \lambda$ and q are hyper-parameters for the prior distribution.

3.1.3 Prior #2 (An Improper Prior)

We considered using improper prior distributions analogous to the distributions used in the context of mixture modelling [25, 15].

We considered using a uniform prior for each σ_j and each μ_j inversely proportional to the standard deviation of the differences:

$$h(\sigma_j) \propto \frac{1}{\sigma_{\Delta y}} \qquad \text{for } \sigma_j \geq 0, \qquad j = 0,\ \ldots,\ C$$

$$h(\mu_j) \propto \frac{1}{\sigma_{\Delta y}} \qquad \text{for } \mu_j \in [-\infty,\ \infty], \qquad j = 0,\ \ldots,\ C$$

We assume that the parameters are independent:

$$h(\mu_j,\ \sigma_j) \propto \frac{1}{\sigma_{\Delta y}^2} = \frac{\rho}{\sigma_{\Delta y}^2} \tag{4}$$

We use $\rho = 1/2$ as this reflects that each σ_j is uniform between 0 and the standard deviation of the differences, and each μ_j is uniform in the region within one standard deviation of the average difference. We find that this prior distribution is scale invariant (i.e., when using Bayesian methods such as Bayes factors or MML we obtain equivalent results independent of whether we measure our time series in dollars or cents).

3.1.4 Prior #3

We assume that each σ_j is a priori distributed according to a Gamma distribution [4, p. 560] $\sigma_j \sim \Gamma(\alpha,\ \beta)$. Again, we assume that we have rough knowledge about the average difference between data points and the average variability in this difference.

A convenient way to make this prior scale invariant is to set $\beta = \sigma_{\Delta y}$. We obtain a reasonable prior distribution if we set $\alpha = 1$ since this doesn't exclude small σ_j and allows unbounded values of σ_j (but with small density):

$$\sigma_j \sim \Gamma(\alpha = 1,\ \beta = \sigma_{\Delta y}) \tag{5}$$

We assume that each μ_j is a priori distributed according to a Gaussian distribution:

$$\mu_j \sim N(\mu_{\Delta y},\ \sigma_{\Delta y}) \tag{6}$$

3.1.5 Discussion of the Prior Distributions

These three prior distributions over σ, along with the traditional $\frac{1}{\sigma}$ prior distribution are shown in Figure 2(a) for the value $\sigma_{\Delta y} = 1$. The $\frac{1}{\sigma}$ prior distribution is normalised to be in the range $[0.01, 2]$.

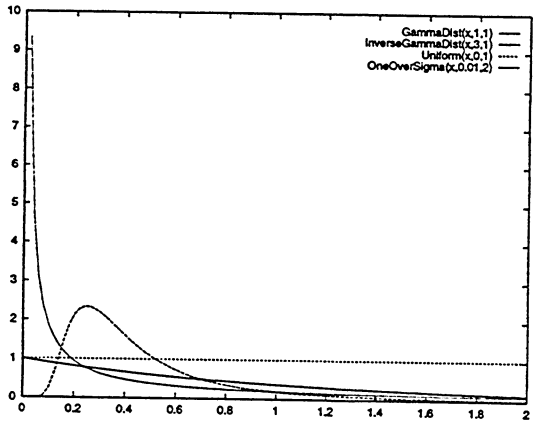

Figure 2: (a) The priors with $\sigma_{\Delta y} = 1$

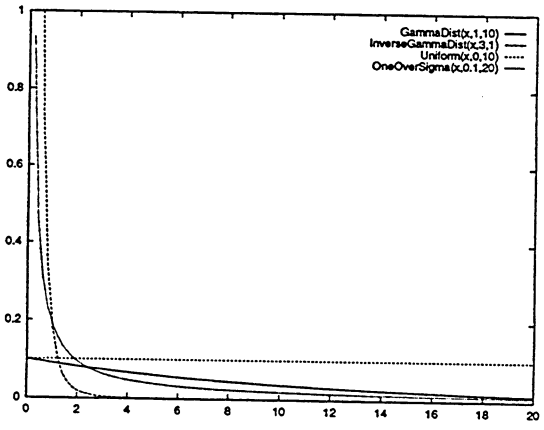

Figure 2: (b) The priors with $\sigma_{\Delta y} = 10$

To see the scale invariance of Prior #2 and Prior #3, Figure 2(b) gives the prior distributions with $\sigma_{\Delta y} = 10$ and renormalises the $\frac{1}{\sigma}$ prior to be in the range $[0.1, 20]$.

3.2 The Bayes Factor or Evidence Approach

Due primarily to the recent development of fast algorithms for approximating high dimensional integrals, interest has been renewed in Bayes factors for use in Bayesian model selection procedures; see Kass and Raftery [8]. In particular, Koop and Potter [9] have used Bayes factors to compare linear and nonlinear models in time series.

The basic setup for comparing hypotheses $H_0, \ldots, H_M$ using Bayes factors is as follows: Let $f_C(x_1 \ldots, x_n \mid \phi_C)$, and $h(\phi_C)$ denote the likelihood function and prior density functions, respectively, for each of the models corresponding to hypothesis H_C under consideration.

For the segmentation problem, it is convenient to partition the hypotheses to correspond to the number of cutpoints:

H_0: there are $C = 0$ cutpoints,

$\vdots$

H_M: there are $C = M$ cutpoints.

Then, the marginal probability density for model H_C is given by

$$m_C(y_1, \ldots, y_N) = \sum_{v \in V_C} h(v) \int f_C(y_1, \ldots, y_N \mid \theta_C,\ v) h(\theta_C) d\theta_C \tag{7}$$

If the hypotheses have equal prior probabilities, then the Bayes factor approach selects the H_C with maximum m_C. Once the hypothesis is chosen, point and interval estimates for parameters based on posterior distributions are formed *conditional* on the chosen hypothesis.

3.3 Calculating the Marginal Density for the Observed Data

To calculate the Bayes factor for segmented models, we first calculate the marginal density of the observed data conditional on the number and location of the cut points. Then, we sum those (conditional) marginal densities having the same number of segments, and divide by the total number of such marginal densities to obtain the unconditional marginal densities. This, of course, assumes the location of cut points are *a priori* equally likely.

3.3.1 The Marginal Density for a Gaussian Sample

To calculate the conditional marginal densities, we suppress the notation regarding the location and number of cutpoints, and calculate the marginal density of the observed data within a segment. Consider $X_1, X_2, \ldots, X_n$ an independent sample from a Normal distribution having mean μ and variance σ^2. We can obtain data in this form by applying the transformation

$$x_n = y_{n+1} - y_n$$

The marginal probability density for the data $x_1, \ldots, x_n$, is calculated by integrating the product of the likelihood function,

$$\begin{aligned} f(x_1, \ldots, x_n \mid \mu, \sigma^2) &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2} \sum_{t=1}^{n} (x_t - \mu)^2\} \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right\}, \end{aligned}$$

and the prior joint density function (for mathematical convenience we used Prior #1):

$$\begin{aligned} h(\mu, \sigma^2) &= h(\mu \mid \sigma^2) h(\sigma^2) \\ &= (2\pi)^{-\frac{1}{2}} \left[\frac{\sigma^2}{q}\right]^{-\frac{1}{2}} \exp\{-\frac{q}{2\sigma^2}(\mu - \lambda)^2\} \frac{1}{\Gamma(\alpha)\beta^\alpha} \sigma^{-2(\alpha+1)} \exp\{-\frac{1}{\sigma^2 \beta}\} \end{aligned}$$

with respect to the unknown parameter variables μ and σ^2.

Integrating first with respect to μ, we obtain

$$\begin{aligned} m_G(x_1, \ldots, x_n \mid \sigma^2) &= \int f(x_1, \ldots, x_n \mid \mu, \sigma^2) h(\mu \mid \sigma^2) d\mu \\ &= \frac{q^{\frac{1}{2}} \sigma^{-(n+1)}}{(2\pi)^{\frac{n+1}{2}}} \exp\{-\frac{(n-1)s^2}{2\sigma^2}\} \int_{-\infty}^{\infty} \exp\{-\frac{[n(\bar{x} - \mu)^2 + q(\mu - \lambda)^2]}{2\sigma^2}\} d\mu \\ &= \frac{q^{\frac{1}{2}} \sigma^{-(n+1)}}{(2\pi)^{\frac{n+1}{2}}} \exp\{-\frac{(n-1)s^2}{2\sigma^2}\} \\ &\quad \cdot \exp\{\frac{(n\bar{x} + q\lambda)^2}{2(n+q)\sigma^2} - \frac{(n\bar{x}^2 + q\lambda^2)}{2\sigma^2}\} \int_{-\infty}^{\infty} \exp\{-\frac{(n+q)}{2\sigma^2}(\mu - \mu*)^2\} d\mu \\ &= \frac{q^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}(n+q)^{\frac{1}{2}}} \sigma^{-n} \exp\{-\frac{1}{\sigma^2}\left[\frac{(n-1)s^2}{2} + \frac{nq(\bar{x} - \lambda)^2}{2(n+q)}\right]\}. \end{aligned}$$

where $\mu* = (n\bar{x} + q\lambda)/(n+q)$.

Multiplying the above with the marginal prior for σ^2, and integrating with respect to σ^2, we obtain the marginal probability density:

$$\begin{aligned} m_G(x_1,\ldots,x_n) &= \int m_G(x_1,\ldots,x_n \mid \sigma^2)h(\sigma^2)d\sigma^2 \\ &= \int_0^\infty \frac{q^{\frac{1}{2}}\sigma^{-2(\frac{n}{2}+\alpha+1)}}{(2\pi)^{\frac{n}{2}}(n+q)^{\frac{1}{2}}\Gamma(\alpha)\beta^\alpha}\exp\{-\frac{1}{\sigma^2}\left[\frac{(n-1)s^2}{2}+\frac{nq(\bar{x}-\lambda)^2}{2(n+q)}+\frac{1}{\beta}\right]\}d\sigma^2 \\ &= \frac{q^{\frac{1}{2}}\Gamma(\frac{n}{2}+\alpha)}{(2\pi)^{\frac{n}{2}}(n+q)^{\frac{1}{2}}\Gamma(\alpha)\beta^\alpha}\left[\frac{(n-1)s^2}{2}+\frac{nq(\bar{x}-\lambda)^2}{2(n+q)}+\frac{1}{\beta}\right]^{-(\frac{n}{2}+\alpha)}. \end{aligned}$$

3.3.2 The marginal probability density for a segmented model

Let v be a segmentation with C cutpoints ($v \in V_C$). Let v segment the $x_1,\ldots,x_{N-1}$ into:

$$\begin{aligned} &x_{01},\ x_{02},\ \ldots,\ x_{0n_0} \\ &x_{11},\ x_{12},\ \ldots,\ x_{1n_1} \\ &\vdots \\ &x_{C1},\ x_{C2},\ \ldots,\ x_{Cn_C} \end{aligned}$$

where n_j is the number of points in segment j.
Assuming that each segment is independent:

$$m(y_1,\ldots,y_N \mid v) = \prod_{j=0}^{C} m_G(x_{j1},\ x_{j2},\ \ldots,\ x_{jn_j})$$

Substituting this equation and Equation (1) into Equation (7) gives us:

$$m_C(y_1,\ldots,y_N) = \frac{\sum_{v\in V_C} m(y_1,\ldots,y_N \mid v)}{|V_C|} \tag{8}$$

which is the quantity we wish to maximise.

3.4 The Minimum Message Length Approach

Wallace et al [25, 26, 27] advocate the use of Minimum Message Length (MML) for Bayesian point estimation. The MML approach can be interpreted as partitioning the parameter space into regions, where the models within a region are considered *similar*. Each region, R, is identified by a representative vector of parameter values θ'_R where $\theta'_R \in R$. We construct a *two-part* message of the form:

- Part 1 specifies a region R, and
- Part 2 describes the data under the assumption that θ'_R is true.

The *essential* aspect of MML which distinguishes it from traditional Bayesian methods is the determination of the size of each region. The partitioning of the parameter space is done in such a way as to minimise the expected length of a message.

An example partitioning (with representative points) might be:

R_{0A}: $C = 0$, $\sigma_0 \in [.4,\ .6]$, $\mu_0 \in [0,\ .2]$.
$\theta'(R_{0A}) = (\sigma_0 = 0.5,\ \mu_0 = 0.1)$
R_{0B}: $C = 0$, $\sigma_0 \in [.4,\ .6]$, $\mu_0 \in [.2,\ .4]$.
$\theta'(R_{0B}) = (\sigma_0 = 0.5,\ \mu_0 = 0.3)$
$\vdots$
R_{0M}: $C = 0$, $\sigma_0 \in [.6,\ .9]$, $\mu_0 \in [0,\ .4]$.
$\theta'(R_{0M}) = (\sigma_0 = 0.75,\ \mu_0 = 0.2)$
R_{0N}: $C = 0$, $\sigma_0 \in [.6,\ .9]$, $\mu_0 \in [.4,\ .8]$.
$\theta'(R_{0M}) = (\sigma_0 = 0.75,\ \mu_0 = 0.6)$
$\vdots$
R_{0X}: $C = 0$, $\sigma_0 \in [.9,\ 1.3]$, $\mu_0 \in [0,\ .5]$.
$\theta'(R_{0X}) = (\sigma_0 = 1.1,\ \mu_0 = 0.25)$
R_{0Y}: $C = 0$, $\sigma_0 \in [.9,\ 1.3]$, $\mu_0 \in [.5,\ 1.0]$.
$\theta'(R_{0Y}) = (\sigma_0 = 1.1,\ \mu_0 = 0.75)$
$\vdots$
R_{1A}: $C = 1$, $v \in [5,\ 8]$, $\sigma_0 \in [.4,\ .6]$, $\mu_0 \in [0,\ .2]$, $\sigma_1 \in [.4,\ .6]$, $\mu_1 \in [.2,\ .4]$.
$\theta'(R_{1A}) = (v = 6.5,\ \sigma_0 = 0.5,\ \mu_0 = 0.1,\ \sigma_1 = 0.5,\ \mu_1 = 0.3)$
R_{1B}: $C = 1$, $v \in [8,\ 12]$, $\sigma_0 \in [.4,\ .6]$, $\mu_0 \in [0,\ .2]$, $\sigma_1 \in [.4,\ .6]$, $\mu_1 \in [.2,\ .4]$.
$\theta'(R_{1B}) = (v = 10,\ \sigma_0 = 0.5,\ \mu_0 = 0.1,\ \sigma_1 = 0.5,\ \mu_1 = 0.3)$
$\vdots$

3.4.1 The Wallace and Freeman MML Approach

Wallace and Freeman [27] used the following method to approximate the message length of y encoded using θ (a vector of continuous parameters).

Under certain regularity conditions, the volume of a region whose representative point is θ' is of size:

$$Vol(\theta') = \frac{\kappa_d^{-\frac{d}{2}}}{\sqrt{det(F(\theta'))}}$$

where d is the dimension of θ, $det(F(\theta'))$ is the determinant of the Fisher Information matrix, and κ_d is the d dimensional optimal quantizing lattice constant (given in Conway and Sloane [6]). The volumes ($Vol(\theta')$) will vary around the parameter space. For example, that part of the parameter space with low σ_j will have a smaller regions than that part with high σ_j (see Section 4.2 of [14]).

We approximate the prior probability of region R (with representative point θ'_R) as:

$$\int_R h(\theta)d\theta \approx Vol(\theta'_R)\ h(\theta'_R)$$

where $h(\theta)$ is the assumed known prior density on θ. The probability of obtaining data y under the assumption that θ'_R is true is:

$$f(y|\theta'_R)\delta_y$$

where $f(y|\theta)$ is the likelihood function and δ_y is the precision to which y was measured.

The probability we associate with y and θ'_R is therefore:

$$Prob(y\&\theta'_R) \;=\; Vol(\theta'_R)h(\theta'_R) \;\times\; f(y|\theta'_R)\delta_y$$

We wish to extend the method to θ which are *not* representative of some region. Doing this leads (on average) to a message which is $d/2$ nits longer. The message length is therefore approximately[1]:

$$MessLen(y\&\theta) \;=\; -\log Prob(y\&\theta) \;+\; \frac{d}{2} \tag{9}$$

3.4.2 Applying MML to Cutpoint-like Parameters

The Wallace and Freeman approach (leading to Equation 9) is not directly applicable to the segmentation problem. Cutpoint-like parameters v do not satisfy the regularity conditions required, and the Fisher Information matrix is not defined for this type of parameter.

Oliver, Baxter and Wallace [16] derive expressions for the width of a region for cutpoint-like parameters:

$$Width(v) \;=\; \frac{4}{(n_0 - n_1)\log\frac{\sigma_1}{\sigma_0} + \frac{RSS_0 - RSS_1 + n_0 D^2}{2\sigma_1^2} + \frac{RSS_1 - RSS_0 + n_1 D^2}{2\sigma_0^2}}$$

where n_0 and n_1 are the number of points in the segment either side of v', σ_0 and σ_1 are the variances of the segments either side of v', RSS_0 and RSS_1 are the residual sum of squares of the segment either side of v', and D is the difference in the means.

Let v'_R be the representative cutpoint for region R. The probability we associate with y, v'_R and θ'_R is therefore:

$$Prob(y\&v'_R\&\theta'_R) \;=\; Vol(\theta'_R)h(\theta'_R) \;\times\; Width(v'_R)h(v'_R) \;\times\; f(y|\theta, v)\delta_y \tag{10}$$

Again, we wish to extend the method to v which are *not* representative of some region. Oliver, Baxter and Wallace [16] show that on average this introduces an additional length of C nits. Hence the message length is approximately

$$MessLen(y\&v\&\theta) \;=\; -\log Prob(y\&v\&\theta) \;+\; \frac{d}{2} \;+\; C \tag{11}$$

3.5 Why is the MML Different from the Bayes Factor Approach?

A number of authors who advocate the Bayes factors (or Evidence) approach have suggested that MDL is equivalent to Bayes factors[5, 12], and that MML is an approximation to Bayes factors[13]:

[1] By coding theory, we can encode an event of probability P in a message of length $-\log P$ nits. A nit is the unit of message length when we take logarithms to the base e. Hence 1 nit $\approx$ 1.44 bits.

"The effectiveness of MML and MDL methods which use two-part codes is because they approximate the log of the evidence. It is more appropriate to approximate the evidence directly."

To examine this issue we give procedures for the two approaches[2][3][4] in Figure 3.

MML inference procedure:

1. Identify ϕ the parameters we wish to estimate.
2. Partition ϕ's parameter space into regions such that using these regions would minimise the expected length of a *two part* message. Associate region l as hypothesis H_l.
3. Select the H_l with minimum message length; this will typically be the model with maximum associated probability by Equation (10).

Bayes Factor inference procedure:

1. Identify the parameter ω we determine as suitable for estimating in the first stage of inference.
2. Partition ω's parameter space into regions. We often set hypothesis H_C to be the models with $\omega = C$.
3. Select the H_C with maximum posterior probability.
4. Estimate the remaining parameters with the constraint that the parameters are in the preferred H_C.

Figure 3: The MML and Bayes Factors Inference Procedures

The MML and Bayes factors procedures have the following differences:

- The MML procedure treats the parameters in a symmetric way, where the Bayes Factor procedure encourages us to distinguish between discrete parameters and continuous parameters.
- The Bayes Factor procedure has the subjective choices of which parameter to select as ω (Step 1), and how to partition ω's parameter space (Step 2).

The subjective choices we made in applying the Bayes factors approach to segmentation were:
(a) $\omega = C$, the number of cutpoints, and
(b) we should partition the hypotheses into:

H_0: there are $C = 0$ cutpoints,
$\vdots$
H_M: there are $C = M$ cutpoints.

[2]The MML inference procedure uses two part messages in Step 2. Rissanen [20] advocates one part messages (which are shorter). The consequence of using one part messages is that the partitioning in Step 2 results in each point in the parameter space having its own partition — no parameter estimation is achieved.

[3]In MacKay's terminology Steps 1-3 of the Bayes Factor procedure are 'Level 2 Inference' and Step 4 is 'Level 1 Inference'.

[4]The estimators used in Step 4 of the Bayes Factor inference procedure may include the mode of the posterior (MAP), the mean of the posterior, etc.

4 Simulations

We use the following criteria:

- AIC, using $-\log f(y|\phi) + d + C$ [11].
- BIC, using $-\log f(y|\phi) + \frac{d+C}{2}\log n$ [10].
- MDL, using $-\log f(y|\phi) + \frac{d}{2}\log n + \log \binom{n}{C}$ [7, 17, 18].
- BF, the Bayes Factor approach described in Section 3.2–3.3, with Prior #1 (with hyper-parameters λ, q, α and β) given in Relations (3) and (3).
- MML1, using Equation (11) of this paper, with Prior #1 (with hyper-parameters λ, q, α and β) given in Relations (3) and (3).
- MML2, using Equation (11) of this paper, with Prior #2 (the uniform improper prior) given in Equation (4).
- MML3, using Equation (11) of this paper, with Prior #3 given in Relations (5) and (6).
- ORAC, an oracle which selects the correct number of segments.

The criteria estimate parameters in different ways. The AIC, BIC, MDL and MML1-3 criteria select the set of parameter values which minimise the criteria. This results in the AIC, BIC and MDL criteria selecting the maximum likelihood estimates for v, μ_j and σ_j^2 from the set of segmentations with the value of C which minimises the criteria. The MML methods do not use the maximum likelihood estimates for v, μ_j and σ_j^2, rather these methods use the estimates for v, μ_j and σ_j^2 which minimise the message length. The BF and ORAC methods first select the number of cutpoints, and then use the maximum likelihood estimates[5] for v, μ_j and σ_j^2.

4.1 Simulation #1

The first simulation involved creating time series data using Prior #1 (from Section 3.1.2). We set the hyper-parameters of the prior distribution to be:

$$q = 0.01 \quad \alpha = 3 \quad \beta = 1 \quad \text{and} \quad \lambda = 0$$

We generated n data points and applied the criteria in Section 4 to estimate the parameters of the distribution.

For the MML1-3, AIC, BIC and MDL criteria we performed an exhaustive search of segmentations where each segment contained 3 or more data items, searching between $k = 1$ and $k = 3$ segments.

For the Bayes Factor criteria (BF) we evaluated the posterior probabilities

$$Prob(k = 1|y),\ Prob(k = 2|y) \text{ and } Prob(k = 3|y)$$

[5]For the BF method this is equivalent to selecting the MAP estimate under a flat prior over v, μ_j and σ_j^2.

	Counts			Mean
	$\hat{k}=1$	$\hat{k}=2$	$\hat{k}=3$	KL
	n=20			
AIC	45	17	38	13.992
BIC	74	15	11	12.841
MDL	92	4	4	9.408
BF	100	0	0	0.063
MML1	100	0	0	0.058
MML2	97	3	0	0.086
MML3	100	0	0	0.058
ORAC	100	0	0	0.063
	n=40			
AIC	17	14	69	14.606
BIC	84	8	8	10.275
MDL	98	2	0	0.070
BF	100	0	0	0.034
MML1	100	0	0	0.032
MML2	100	0	0	0.032
MML3	100	0	0	0.032
ORAC	100	0	0	0.034
	n=80			
AIC	5	6	89	11.129
BIC	94	2	4	3.082
MDL	100	0	0	0.013
BF	100	0	0	0.013
MML1	100	0	0	0.013
MML2	99	0	1	0.015
MML3	100	0	0	0.013
ORAC	100	0	0	0.013

Table 1: (a) True no. of segments = 1

	Counts			Mean
	$\hat{k}=1$	$\hat{k}=2$	$\hat{k}=3$	KL
	n=20			
AIC	0	38	62	4.094
BIC	3	63	34	3.686
MDL	3	83	14	3.267
BF	6	93	1	0.244
MML1	9	91	0	0.134
MML2	5	94	1	0.126
MML3	8	92	0	0.130
ORAC	0	100	0	0.182
	n=40			
AIC	1	30	69	3.326
BIC	4	74	22	2.742
MDL	6	83	11	2.597
BF	7	91	2	0.084
MML1	7	92	1	0.072
MML2	6	91	3	0.070
MML3	7	91	2	0.070
ORAC	0	100	0	0.080
	n=80			
AIC	0	25	75	0.741
BIC	2	88	10	0.180
MDL	2	94	4	0.094
BF	2	97	1	0.043
MML1	3	95	2	0.043
MML2	2	96	2	0.040
MML3	2	97	1	0.040
ORAC	0	100	0	0.040

Table 1: (b) True no. of segments = 2

Table 1 lists the number of times the criteria estimated each value of k from 100 simulations. In addition, Table 1 gives the average Kullback-Liebler distance (Mean KL) between the predicted distribution, and the underlying distribution [6].

[6]The Kullback-Liebler distance (given for example in [23, Chp. 9]) between a true distribution $N(\mu_t,\ \sigma_t^2)$ and a fitted distribution $N(\mu_f,\ \sigma_f^2)$ is

$$\log\frac{\sigma_f}{\sigma_t} - \frac{1}{2} + \frac{1}{2\sigma_f^2}(\sigma_t^2 + (\mu_t - \mu_f)^2).$$

4.2 Simulation #2

4.2.1 The Search Method

It is impractical to consider every possible segmentation of data once we consider multiple cutpoints. We therefore used the following search method. Given a set of data, we consider every binary segmentation (i.e., one cutpoint) and identify those cutpoints which are local maxima in likelihood. We then perform an exhaustive search of segmentations using the cutpoints which are local maxima in likelihood. The segmentations are also required to have a minimum segment length of 3.

4.2.2 Results

True no. of segments = 1						
	$\hat{k}$					Mean
	1	2	3	4	5	KL
	n=20					
AIC	48	31	17	4	0	15.620
BIC	79	16	5	0	0	14.646
MDL	94	5	1	0	0	14.101
MML2	100	0	0	0	0	0.065
MML3	100	0	0	0	0	0.065
ORAC	100	0	0	0	0	0.071
	n=40					
AIC	27	17	31	19	6	8.969
BIC	80	15	5	0	0	6.199
MDL	94	6	0	0	0	2.216
MML2	100	0	0	0	0	0.028
MML3	100	0	0	0	0	0.028
ORAC	100	0	0	0	0	0.029
	n=80					
AIC	15	11	35	16	23	2.736
BIC	93	7	0	0	0	0.675
MDL	99	1	0	0	0	0.267
MML2	100	0	0	0	0	0.013
MML3	100	0	0	0	0	0.013
ORAC	100	0	0	0	0	0.013
	n=160					
AIC	3	12	41	27	17	4.822
BIC	94	4	2	0	0	1.580
MDL	100	0	0	0	0	0.006
MML2	97	3	0	0	0	0.008
MML3	99	1	0	0	0	0.007
ORAC	100	0	0	0	0	0.006

Table 2: (a) True no. of segments = 1

True no. of segments = 2						
	$\hat{k}$					Mean
	1	2	3	4	5	KL
	n=20					
AIC	19	47	30	4	0	17.882
BIC	40	44	16	0	0	17.330
MDL	66	26	8	0	0	14.757
MML2	88	11	1	0	0	0.200
MML3	98	2	0	0	0	0.171
ORAC	0	100	0	0	0	7.129
	n=40					
AIC	3	34	35	21	7	15.025
BIC	30	51	16	3	0	14.138
MDL	59	34	6	1	0	12.336
MML2	56	38	6	0	0	0.159
MML3	75	24	1	0	0	0.145
ORAC	0	100	0	0	0	2.200
	n=80					
AIC	0	25	34	26	15	5.322
BIC	15	79	6	0	0	2.510
MDL	30	66	4	0	0	2.397
MML2	22	72	5	1	0	0.075
MML3	33	66	1	0	0	0.078
ORAC	0	100	0	0	0	0.279
	n=160					
AIC	0	18	33	34	15	3.940
BIC	1	95	4	0	0	2.263
MDL	3	95	2	0	0	2.182
MML2	1	94	5	0	0	0.026
MML3	2	97	1	0	0	0.026
ORAC	0	100	0	0	0	0.024

Table 2: (b) True no. of segments = 2

The data used in Simulation #2 were generated according to the following distributions:

- Table 2(a) — One segment generated by the time series $y_{t+1} = y_t + \epsilon_t$.
- Table 2(b) — Two segments with the first half generated by $y_{t+1} = y_t + \epsilon_t$, and the second half generated by $y_{t+1} = y_t + 1 + \epsilon_t$.
- Table 3 — Three segments with the first third generated by $y_{t+1} = y_t + \epsilon_t$, the middle third generated by $y_{t+1} = y_t + 1 + \epsilon_t$ and the last third generated by $y_{t+1} = y_t + 2 + \epsilon_t$.

In all cases the ϵ_t were generated from a Gaussian with mean 0 and variance 1, $N(0,1)$.

	$\hat{k}$						Mean
	1	2	3	4	5	6	KL
	n=20						
AIC	2	51	37	9	1	0	16.393
BIC	11	66	20	3	0	0	16.014
MDL	28	57	13	2	0	0	15.621
MML2	55	42	3	0	0	0	0.340
MML3	77	23	0	0	0	0	0.338
ORAC	0	0	100	0	0	0	16.279
	n=40						
AIC	0	10	52	34	4	0	13.727
BIC	1	73	23	3	0	0	11.733
MDL	7	77	15	1	0	0	11.603
MML2	3	67	29	1	0	0	0.192
MML3	21	74	5	0	0	0	0.199
ORAC	0	0	100	0	0	0	12.299

	$\hat{k}$						Mean
	1	2	3	4	5	6	KL
	n=80						
AIC	0	3	33	42	16	6	2.657
BIC	0	61	34	5	0	0	1.956
MDL	1	82	14	3	0	0	1.201
MML2	0	42	58	0	0	0	0.099
MML3	1	58	41	0	0	0	0.110
ORAC	0	0	100	0	0	0	1.496
	n=160						
AIC	0	0	39	33	20	8	2.813
BIC	0	22	75	3	0	0	2.309
MDL	0	39	60	1	0	0	2.221
MML2	0	7	89	4	0	0	0.047
MML3	0	13	87	0	0	0	0.048
ORAC	0	0	100	0	0	0	2.229

Table 3: True no. of segments = 3

In each simulation, we generated n points from the time series. We applied the search method described in Section 4.2.1 searching between $k = 1$ and $k = 5$ or $(k = 6)$ segments. We applied the AIC BIC, MDL, MML2-3 and ORAC criteria during this search and identified a preferred segmentation for each of these criteria. We did not use BF in this simulation since (a) the search method used does not consider all the segmentations we are required to sum in Equation (8), and (b) it is not clear how we would set the hyper-parameters for this problem. Tables 2 and 3 list the number of times the criteria estimated each value of k from 100 simulations, and give the average Kullback-Liebler distance (Mean KL) between the fitted distribution, and the true distribution.

5 Discussion of Results

5.1 Discussion of Simulation #1

The results in Table 1 indicate that the Bayesian approaches (namely MML1, MML2, MML3 and BF) were superior. This was to be expected for two of these approaches (MML1 and BF),

since they had access to the values of the hyper-parameters used to generate the data. The MML approaches using 'ignorance' prior distributions, MML2 and MML3, performed exceptionally well since they had access to the same information as AIC, BIC and MDL (they didn't have access to the hyper-parameters' values).

5.2 Discussion of Simulation #2

There are three notable features of the results in Tables 2 and 3:

1. The AIC criterion appears to be inappropriate for this task.

2. The other criteria (BIC, MDL and MML2-3) appear to identify the number of segments adequately.

3. Even though the ORAC criterion has access to the correct number of cutpoints, it gets very poor mean Kullback-Liebler distances. The mean Kullback-Liebler distances for the BIC and MDL criteria are also quite poor. We discuss this issue in the next section.

We provided results for MML using two prior distributions. The prior distributions used by the MML method had an effect on the choice of the number of segments. However, the MML method consistently selected segmentations with low Kullback-Liebler distances with both prior distributions used.

5.3 Issues about the Kullback-Liebler Distance

The MML methods outperform the oracle (ORAC) method in mean Kullback-Liebler distance. This is a surprising result as the oracle (ORAC) method 'knows' the number of segments used to generate the data. We suggest that the ORAC criterion is getting poor mean Kullback-Liebler distances because the Maximum Likelihood estimates of v are unreliable.

To confirm this hypothesis, we examined the simulations from Table 2(b) when $n = 40$. For these 100 simulations, we examined the Maximum Likelihood and MML estimates for v. Figure 4 gives histograms for the location of the selected cutpoints (out of the 39 possible positions). The MML estimates for v are clearly closer to the correct cutpoint ($v = 19$). Segmentations with cutpoints well away from $v = 19$ will have significantly higher Kullback-Liebler distances.

5.4 The Kullback-Liebler Distance of the Bayes Factor Approach

The Bayes Factor approach may suffer from the problem of selecting inappropriate cutpoints. If we use flat prior distributions over the v, μ_j and σ_j. then the MAP estimate for the cutpoint will be the same as the Maximum Likelihood estimate. The ORAC criterion is equivalent to an optimal Bayes Factor approach with flat prior distributions (such as Prior #2) and using the MAP estimate for v, μ_j and σ_j.

If one uses a two stage Bayes factors procedure, then it is unclear to us as to how to correct the problem of estimating the v, μ_j and σ_j in a 'traditional' Bayesian manner:

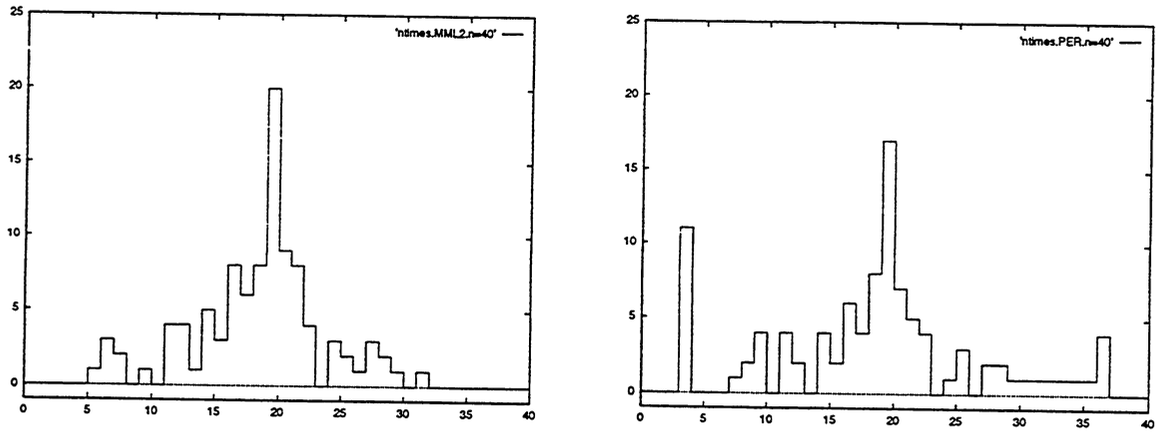

Figure 4: (a) MML estimates of v (b) Maximum Likelihood estimates of v

- The MAP estimate will be the same as the Maximum Likelihood estimate for flat prior distributions. The Maximum Likelihood estimate will often produce models with high Kullback-Liebler distances.

- The mean or median of the posterior may be sensible estimators for the $C = 1$ case. To do this, we might use the mean or median of the posterior over v to select v, and then estimate μ_j and σ_j conditional on v. This approach appears difficult to apply to the $C > 1$ case.

6 Applications

6.1 The US GNP 1947 – 1966

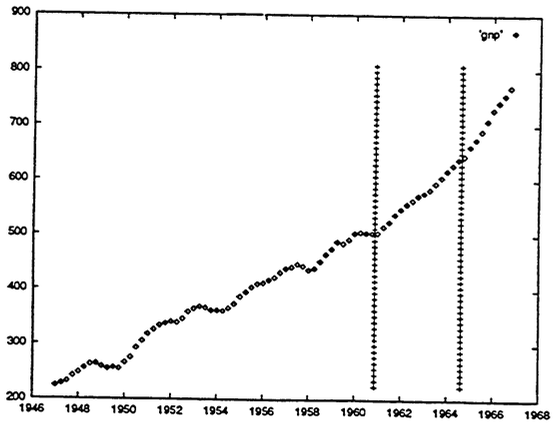

Figure 5: The US GNP 1947 – 1966

We segmented the quarterly gross national product (GNP) for the United States from 1947

– 1966 [22]. Figure 5[7] shows the preferred MML2 segmentation for this data. The BIC and MDL criteria also preferred this segmentation, while the AIC criterion preferred a segmentation with 7 segments.

6.2 The Canadian 10 year Bond Yield 1989 – 1996

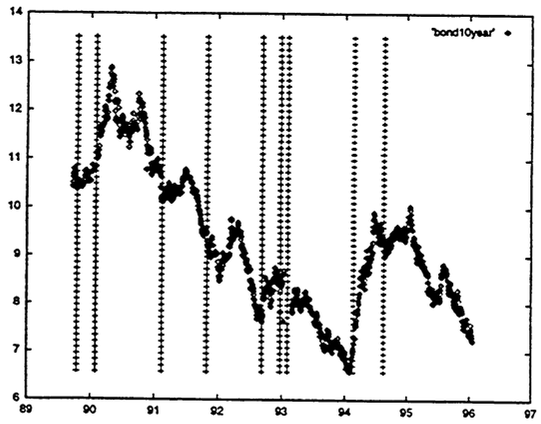

Figure 6: The Canadian 10 year bond yield 1989 – 1996 with 8 cut points

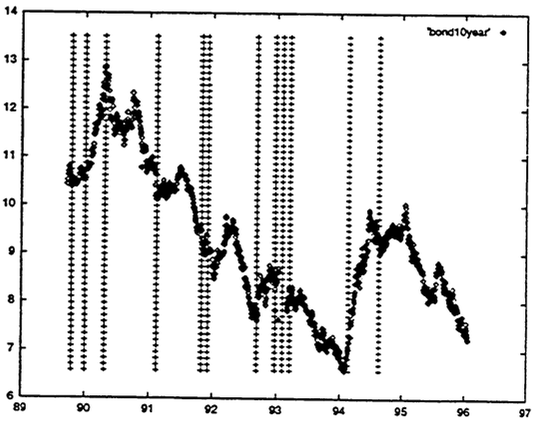

Figure 7: The Canadian 10 year bond yield 1989 – 1996 with 12 cut points

We then considered segmenting a larger data set, namely the Canadian 10 year bond yield. The data set consists of 1514 values of the Canadian 10 year bond (measured in Canadian dollars) for the period 1989 – 1996. The segmentation program took 24 minutes and 31 seconds to examine segmentations of up to 30 segments on a DECstation 5000/20 using a greedy search strategy. The MML2 criterion found evidence for there being at least 8 cut points (see Figure 6) since the message length of the data with no cut points was 5501.9 nits and the message length with 8 cut points was 5295.1 nits. The minimum message length (with 12 cut points – see Figure 7) was 5282.8 nits.

[7]The units in the figure are billions of (non constant) dollars.

7 Conclusion

This paper compared two Bayesian approaches to the segmentation of time series, namely the Bayes Factor (or Evidence) approach and the Minimum Message Length (MML) approach. If one is genuinely *only* interested in estimating how many segments are inherent in a time series, then the Bayes factors approach is preferred. However, if one is also interested in estimating more parameters (e.g, *when* the segments start and finish), then the MML approach was more suited to the segmentation problem considered. The reasons for this are:

- Firstly, the Bayes Factor approach does not offer an adequate method for selecting the position of cutpoints. The Maximum Likelihood estimate for the position of cutpoints is shown to have high Kullback-Liebler distances.
- Secondly, the Bayes Factor approach encourages the use of prior distributions which are mathematically convenient. Mathematically convenient prior distributions may not reflect out prior beliefs, or may have hyper-parameters which are difficult to set.
- Thirdly, the MML method uses approximations which involve less computation than the Bayes Factor approach.

Furthermore, our preferred segmentation technique (MML with data-based prior distributions) significantly outperformed the classical approaches in the simulations we performed.

Acknowledgments

We would like to thank Rohan Baxter, Chris Wallace, Rodney Strachan and David Albrecht for valuable discussions. Jon Oliver acknowledges research support by Australian Research Council (ARC) Postdoctoral Research Fellowship F39340111. Catherine Forbes acknowledges research support by a Monash University Faculty Research Grant.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[2] J. Albert and S. Chib. Bayesian Inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics,* 11, pages 1-15, 1993.

[3] R.A. Baxter and J.J. Oliver. The kindest cut: minimum message length segmentation. In S. Arikawa and A. Sharma, editors, *Lecture Notes in Artificial Intelligence 1160, Algorithmic Learning Theory, ALT-96*, pages 83–90. Springer-Verlag, Berlin, 1996.

[4] J.O. Berger. *Statistical Decision Theory and Bayesian analysis.* Springer-Verlag, 1993.

[5] P. Cheeseman. Personal communication, 1993.

[6] J.H. Conway and N.J.A Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, London, 1988.

[7] B. Dom. MDL estimation with Small Sample Sizes including an application to the problem of segmenting binary strings using bernoulli models. Technical Report RJ 9997 (89085) 12/15/95, IBM Research Division, Almaden Research Center, 650 Harry Rd, San Jose, CA, 95120-6099, 1995.

[8] R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[9] G. Koop and S.M. Potter. Bayes Factors and nonlinearity: Evidence from economic time series. UCLA Working Paper, August 1995, submitted to *Journal of Econometrics.*

[10] Mengxiang Li. Minimum description length based 2-D shape description. In *IEEE 4th Int. Conf. on Computer Vision*, pages 512–517, May 1992.

[11] Z. Liang, R.J. Jaszczak, and R.E. Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with applications to medical image processing. *IEEE Trans. on Nuclear Science*, 39(4):1126–1133, 1992.

[12] David J.C. MacKay. *Bayesian Modeling and Neural Networks.* PhD thesis, Dept. of Computation and Neural Systems, CalTech, 1992.

[13] David J.C. MacKay. Personal communication, 1997.

[14] J.J. Oliver and R.A. Baxter. MML and Bayesianism: Similarities and differences. Technical report TR 206, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[15] J.J. Oliver, Baxter R.A., and Wallace C.S. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 364–372, 1996.

[16] J.J. Oliver, Baxter R.A., and Wallace C.S. Minimum message length segmentation. In *submitted to Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-98)*, 1997. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.

[17] B. Pfahringer. Compression-based discretization of continuous attributes. In *Machine Learning: Proceedings of the Twelfth International Workshop*, pages 456–463, 1995.

[18] J.R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence*, 4:77–90, 1996.

[19] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

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

[21] G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[22] S.L. Sclove. On segmentation of time series. In S. Karlin, T. Amemiya, and L. Goodman, editors, *Studies in econometrics, time series, and multivariate statistics*, pages 311–330. Academic Press, 1983.

[23] C.W. Therrien. *Decision, estimation, and classification : an introduction to pattern recognition and related topics.* Wiley, New York, 1989.

[24] H. Tong. *Non-linear time series : a dynamical system approach.* Clarendon Press, Oxford, 1990.

[25] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.

[26] C.S. Wallace and D.M. Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.

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

[28] A. Zellner. *An introduction to Bayesian inference in economics.* Wiley, New York, 1971.