Minimum Message Length Segmentation

Jonathan J. Oliver[1], Rohan A. Baxter[2] and Chris S. Wallace[1]
jono@ultimode.com, rohan@ultimode.com, csw@cs.monash.edu.au

[1] Dept. Computer Science, Monash University, Clayton Vic., Australia
[2] Ultimode Systems, 2560 Bancroft Way #213, Berkeley, CA 94704, USA

Abstract. The segmentation problem arises in many applications in data mining, A.I. and statistics, including segmenting time series, decision tree algorithms and image processing. In this paper, we consider a range of criteria which may be applied to determine if some data should be segmented into two or regions. We develop a information theoretic criterion (MML) for the segmentation of univariate data with Gaussian errors. We perform simulations comparing segmentation methods (MML, AIC, MDL and BIC) and conclude that the MML criterion is the preferred criterion. We then apply the segmentation method to financial time series data.

1 Introduction

We consider a particular instance of the segmentation problem. The segmentation problem arises wherever it is desired to partition data into distinct homogeneous segments (or regions). The segmentation problem is to decide whether to divide a segment into one or more sub-segments and to choose where to make the divisions.

The segmentation problem arises in applications that partition data in areas such as data mining, A.I. and statistics. The segmentation problem arise in applications such as segmenting time series [14, 16, 5], decision tree algorithms [11, 10], and image processing [7, 6].

1.1 The Problem Considered

Here, we consider a univariate problem, where the segment boundarys are defined by *cut-points*. We assume that the data in each segment is defined by a Gaussian distribution. Figure 1 gives an example of the type of data we might consider. We could ask questions such as “Does this data consist of 1, 2 or 3 segments?”; “If it consists of 3 segments, is the behaviour in the first third the same as the behaviour in the last third?” This paper investigates methods for determining for some data:

(i) how many cut-points should we fit (if any at all)
(ii) the location of the cut-points, and
(iii) estimating the parameters (means and variances) for each segment.

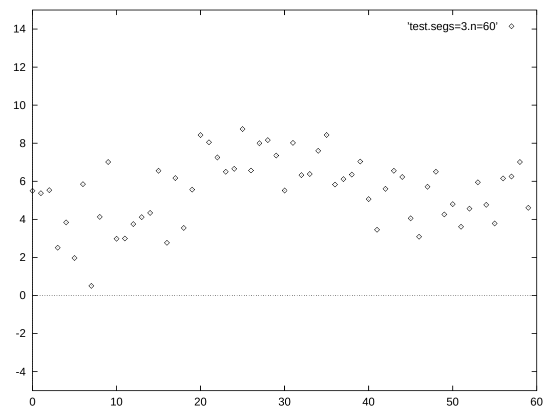

Fig. 1. Example Data for Segmentation

1.2 Motivating the Problem Considered

At first it would appear that the problem as given is overly simple — it would not describe any real world situations, and it should be easy to solve. We argue that these objections are false. Data such as that in Figure 1 might be the number of eye movements per 5 second intervals for a sleeping person, and a doctor may be interested in how many phases of sleep there were, and when they were [14].

A different practical example where this model seems plausible is the incidence of tooth cavities. Previously dentists entertained the burst-remission theory, and dentists spent considerable effort looking for factors that induced remission (i.e., segments with lower means). However, it appears that the data was consistent with the assumption that it was a random walk (i.e. that there was only one segment).

Tong [16] has written a comprehensive book about non linear time series (including segmentation models). We consider such problems in Section 6.

1.3 Related Work

The fit of a segmentation model to data can be expressed precisely using maximum likelihood estimation. However, choosing a segmentation model to maximize the likelihood results in a model with homogeneous regions containing only one datum each. Therefore, heuristics for solving the segmentation problem usually involve 'penalizing' a segmentation for its model complexity. A number of methods which penalize model complexity are available including AIC [1, 7], BIC [13, 6], Minimum Description Length (MDL) [12] and Minimum Message Length (MML) [17, 18].

In this paper, we extend the MML approach to segmentation offered by Baxter and Oliver [2], to the multiple cutpoint case, and apply the approach to time series problems.

This paper is organised as follows: Section 2 defines the segmentation problem we address here. Section 3 describes a previous MDL approach [4, 10, 11], and describes a shortcoming of this approach. Section 4 gives an MML approach to segmentation. The MML method proposed here differs from the MDL approach by optimising the code for the region boundary *and* including coding penalties for stating the parameters of each region. We then compare a variety of segmentation methods on simulations in Section 5. Section 6 applies the method developed to financial time series problem.

2 Notation

Consider some data given as follows. We have n data points, each of which consists of a pair (x_i, y_i). The x_i are evenly spaced between $[0, R]$. The range $[0, R]$ can be cut into $C + 1$ pieces by C segment boundaries (or cutpoints), $\{v_1, \ldots v_C\}$. Each y_i in segment j is distributed with a Gaussian distribution with mean c_j, and standard deviation σ_j.

We wish to estimate the following parameters: (i) C, the number of cutpoints, (ii) the segment boundaries, $\{v_1, \ldots v_C\}$, (iii) the means, $\{c_0, \ldots, c_C\}$, and (iv) the standard deviations, $\{\sigma_0, \ldots, \sigma_C\}$.

3 The Straightforward MDL Approach

Rissanen [12] proposed the straight forward Minimum Description Length (MDL) criterion, which given data y and parameters θ approximates the length as:

$$DescriptionLength(y, \theta) = -\log f(y|\theta) + \frac{number\ params}{2} \log n$$

where $f(y|\theta)$ is the Gaussian likelihood function, $-\log f(y|\theta)$ approximates the length of describing the data, and $\frac{number\ params}{2} \log n$ approximates the length of describing the parameter vector. This approximation is unsuited to cutpoint-like parameters. A number of authors [4, 10, 11] have given terms[3] to describe the cost of stating a cutpoint in a message. A straightforward method of coding a cutpoint is to assume that the cutpoint is equally likely to occur in between x_i and x_{i+1} for $i = 1 \ldots (n-1)$ which leads to a cost[4] of $\log(n)$ to describe the cutpoint. If we wish to state C cutpoints, then this will require a codeword of length:

$$DescriptionLength(C\ cutpoints) = \log \binom{n}{C}$$

Dom [4] requires that $C < \frac{n}{2}$, otherwise the complexity of the term decreases for increasing C, which is counter to prior beliefs about segmentation models in most applications.

[3] We note that these authors used this penalty measure in different, but related contexts and that our use of it here is not meant to imply that these authors would advocate its use here.

[4] Most authors simplify matters by allowing the cutpoint to take n possible values rather than $n - 1$ values.

3.1 A Problem with the Straightforward Approach

A problem with the straightforward MDL approach is that we may use too many bits to describe a cutpoint exactly. Consider a situation where we have the following 17 data points, with points 1-9 been generated by the Gaussian distribution $N(\mu = 0.0,\ \sigma^2 = 1.0)$ and points 10-17 been generated by $N(\mu = 1.0,\ \sigma^2 = 1.0)$:

1	2	3	4	5	6	7	8	9
2.01	-1.78	-1.16	-2.00	-1.68	0.28	0.17	-0.50	0.06

10	11	12	13	14	15	16	17
1.29	-0.43	1.70	0.74	2.69	3.75	0.81	0.66

The straightforward MDL approach requires 4 binary bits to describe a cutpoint, The negative log-likelihood $-\log f(y|\theta)$ is minimised if we place the cutpoint between points 11 and 12. Placing the cutpoint here, results in the following estimates:

$$c_0 = -0.34, \quad c_1 = 1.72, \quad \sigma_0 = 1.22, \quad \sigma_1 = 1.15$$

and a negative log-likelihood: $-\log f(y|\theta) \ = \ 25.68 + 13.50 \ = \ 39.18$ bits. The total description length is then:

$$DescriptionLength(y, \theta) \ = \ 39.18 + 8.17 + 4.00 \ = \ 51.35 \text{ bits}$$

We should also consider encoding the cutpoints less precisely. For example, we could use an encoding scheme which restricts the cutpoints to every second interval, thus requiring only 3 bits to specify a cutpoint. Using this scheme, and placing the cutpoint between points 8 and 9 results in a description length of 40.28 + 8.17 + 3.00 = 51.45 bits.

We can further restrict the possible cutpoints to every fourth interval, thus requiring only 2 bits to specify the cutpoint. Using this scheme, and placing the cutpoint between points 8 and 9 results in a description length of 40.28 + 8.17 + 2.00 = 50.45 bits.

Obviously there are many such schemes — the issue we raise is that we may consider schemes where less that 4 bits are required to encode a cutpoint. However, using fewer bits to describe the cutpoint means that our model is less likely to fit the data well.

The MML approach requires us to determine how precisely we wish to state parameters, and hence the mathematics in this paper optimises the choice of coding schemes for cutpoints.

4 Applying MML to Segmentation

We consider sending a message for this data of the form:

$$C, \quad c_0, \ \ldots \ c_C, \quad \sigma_0, \ \ldots \ \sigma_C, \quad v_1, \ \ldots \ v_C, \quad y_1, \ \ldots \ y_n.$$

The distance between successive x_i is assumed known. Since the x_i are evenly spaced, one can work out the number of x_i in any region from knowing the size of the region. The range of x_i is assumed to be known by the receiver *a priori*.

4.1 Minimum Message Length Formulas

Wallace and Freeman [18] showed that under some fairly general conditions (a locally flat prior and quadratic log-likelihood function) the expected message length (taking the expectation over coding schemes [8, Section 3.3.1]) for sending y and parameters θ is:

$$E(MessLen(y,\theta)) = -\log h(\theta) - \log f(y|\theta) + 0.5 \log det(F(\theta)) + \frac{d}{2}\log \kappa_d + \frac{d}{2}$$

where $h(\theta)$ is the assumed known prior density on θ, d is the dimension of θ, $f(y|\theta)$ is the likelihood, of y given θ, $det(F(\theta))$ is the determinant of the Fisher Information matrix, and κ_d is the d dimensional lattice constant.

The Wallace and Freeman approximation does not apply to cutpoint-like parameters because the log-likelihood function is not continuous, and hence the Fisher Information matrix is not defined for this type of parameter.

4.2 The One Segment, $C = 0$, case

For fitting a constant with no cut points $C = 0$, our θ consists of two parameters, c_0 and σ_0. We choose a non-informative (improper) prior based on the population variance of y_i [17, 9]:

$$h(c_0,\ \sigma_0) = \frac{1}{2\sigma_{pop}^2}$$

where σ_{pop} is the standard deviation of the y_i.

Since the likelihood is Gaussian $N(c_0,\ \sigma_0^2)$, the Fisher Information matrix in this case has two diagonal entries and is:

$$det(F(c_0,\ \sigma_0)) = \frac{2n^2}{\sigma_0^4}$$

For a Gaussian likelihood, the negative log-likelihood, L_0 simplifies:

$$L_0 = -\log f(y|\theta) = n\log(\sqrt{2\pi}\sigma_0) + \sum_{i=1}^{n}\frac{(y_i - c_0)^2}{2\sigma_0^2} = n\log(\sqrt{2\pi}\sigma_0) + \frac{n}{2} \quad (1)$$

Hence, we get the following expression for the expected message length:

$$E(MessLen) = -\log h(c_0, \sigma_0) + 0.5\log det(F(c_0, \sigma_0)) + n\log(\sqrt{2\pi}\sigma_0) + \frac{n}{2} + \frac{\log \kappa_2}{2} + \frac{d}{2}$$

where $d = 2$ and $\kappa_2 = \frac{5}{36\sqrt{3}}$ [3].

4.3 The $C = 1$ case

We now consider the effect of stating the cut point, v, imprecisely. Let the cut point have precision $AOPV_v$ (an acronym for Accuracy Of Parameter Value).

Let ϵ be the difference in the v stated in the message, and the maximum likelihood v estimated from the data. Assume ϵ is uniformly distributed in the range $[\frac{-AOPV_v}{2}, \frac{AOPV_v}{2}]$. We now need to state c_0 and c_1, the constants fitted to the data in the regions on each side of the cut point and also the cut point itself.

In the following we denote the set of x_i in region 0 fitted by constant c_0 as S_0. We do the same for the set of x_i in region 1 fitted by constant c_1, denoting it S_1. Let n_0 be the number of items in S_0 and n_1 be the number of items in S_1. The residual errors are assumed to be distributed as $N(0, \sigma_0^2)$ for region S_0 and as $N(0, \sigma_1^2)$ for region S_1. We assume that the v is uniformly distributed, and hence $h(v) = \frac{1}{R}$. The message length expression for the parameters is then written as follows:

$$\begin{aligned} MessLen(\theta) = & -\log h(c_0,\ \sigma_0) - \log h(c_1,\ \sigma_1) - \log 1/R \\ & +0.5 \log det(F(c_0,\ \sigma_0)) + 0.5 \log det(F(c_1,\ \sigma_1)) - \log AOPV_v \\ & +2 + 2\log \kappa_4 \end{aligned} \tag{2}$$

We note that, given our assumptions about evenly spaced x, we expect $n(1 - \frac{|\epsilon|}{R})$ data items will lie in their correct regions, but we expect $\frac{n|\epsilon|}{R}$ data items will be put in the 'wrong' region.

Let MLC_j be the per item data cost of stating an item *correctly* put in segment j. Hence,

$$MLC_0 \;=\; \log(\sqrt{2\pi}\sigma_0) \;+\; \sum_{i \in S_0} \frac{(y_i - c_0)^2}{2\sigma_0^2 n_0}$$

Let MLW_j be the per item data cost of stating an item *wrongly* put in segment j and hence,

$$MLW_0 \;=\; \log(\sqrt{2\pi}\sigma_1) \;+\; \sum_{i \in S_0} \frac{(y_i - c_1)^2}{2\sigma_1^2 n_0}$$

The message length expression for the data is then:

$$MessLen(y|\theta) = MessLen(y \in \text{correct region}|\theta) \;+\; MessLen(y \in \text{wrong region}|\theta)$$

which we approximate as:

$$\begin{aligned} MessLen(y|\theta) \approx n_0 MLC_0 \;+\; n_1 MLC_1 \;-\; \frac{MLC_0 + MLC_1}{2}\left(\frac{n|\epsilon|}{R}\right) \;+ \\ \frac{MLW_0 + MLW_1}{2}\left(\frac{n|\epsilon|}{R}\right) \end{aligned}$$

We wish to determine the expected message length. The expected cost of stating incorrectly identified data is simplified by letting $D = c_0 - c_1$:

$$E(MLW_0) \;=\; \log(\sqrt{2\pi}\sigma_1) \;+\; \frac{RSS_0 + n_0\ D^2}{2\sigma_1^2 n_0}$$

where RSS_0 is the residual sum of squares ($RSS_0 = \sum_{i \in S_0} (y_i - c_0)^2$).

The expected value of the absolute value of ϵ is $\frac{AOPV_v}{4}$, since

$$E(|\epsilon|) = \frac{2}{AOPV_v} \int_0^{\frac{AOPV}{2}} x \, dx = \frac{AOPV_v}{4}.$$

Hence, the expected message length for the data is:

$$E(MessLen(y|\theta)) = L_0 + L_1 + \left(\frac{nAOPV_v}{8R}\right) E\left(MLW_0 - MLC_0 + MLW_1 - MLC_1\right) \tag{3}$$

where L_0 and L_1 are the negative log likelihoods of segment 0 and segment 1 respectively (as defined in Equation (1)).

We now sum the terms which contain $AOPV_v$ from Equations (2) and (3):

$$-\log AOPV_v + \left(\frac{nAOPV_v}{8R}\right) E\left(MLW_0 - MLC_0 + MLW_1 - MLC_1\right) \tag{4}$$

We take the partial derivative of Expression (4) w.r.t. $AOPV_v$, set the result to 0 and solve for the optimal $AOPV_v$ to minimize the expected message length expression:

$$AOPV_v = \frac{8R/n}{E(MLW_0 - MLC_0 + MLW_1 - MLC_1)}$$

The $AOPV_v$ can be interpreted as a volume in the parameter space. As n_0 and n_1 grow, we see that the volume decreases because the estimate of v can be stated more accurately.

4.4 Message Length Expression

To simplify the algebra, let

$$X = E(MLW_0 - MLC_0) + E(MLW_1 - MLC_1),$$

so that the optimal $AOPV_v$ is $\frac{8R/n}{X}$. We substitute the optimal $AOPV_v$ into the message length expression obtained by summing Equations (2) and (3) and simplifying:

$$\begin{aligned} E(MessLen(y,\theta)) = & -\log h(c_0,\ \sigma_0) - \log h(c_1,\ \sigma_1) - \log 1/R \\ & +0.5 \log det(F(c_0,\ \sigma_0)) + 0.5 \log det(F(c_1,\ \sigma_1)) - \log AOPV_v \\ & +2 + 2\log \kappa_4 + L_0 \ + \ L_1 \ + \ \frac{X}{X} \end{aligned} \tag{5}$$

4.5 Multiple Cutpoints

We now generalise Equation (5) to $C > 1$ cutpoints. Let MLC_j be the per item data cost of stating an item *correctly* put in segment j. Let $MLW_{j,k}$ be the per item data cost of stating an item from segment j *wrongly* put into segment k. For each cutpoint ($j = 1..C$) let

$$X_j = E(MLW_{j-1,j} - MLC_{j-1}) + E(MLW_{j,j-1} - MLC_j)$$

so that the optimal $AOPV_{vj}$ for cutpoint j is:

$$AOPV_{vj} \;=\; \frac{8R/(n_{j-1} + n_j)}{X_j}$$

With $C > 1$ cutpoints, we have:

$$E(MessLen(y, \theta)) = -\sum_{j=0}^{C} \log h(c_j,\ \sigma_j) \;-\; C \log 1/R \;+\; 0.5 \sum_{j=0}^{C} \log det(F(c_j,\ \sigma_j))$$
$$- \log C! - \sum_{j=1}^{C} \log AOPV_{vj} + \frac{d}{2} + \frac{d}{2} \log \kappa_d + \sum_{j=0}^{C} L_j + C \qquad (6)$$

5 Simulation Results

We ran simulations comparing the following criteria:
(i) MML, using Equation (6) of this paper.
(ii) AIC, using $-\log f(y|\theta) + number\ params$ [7].
(iii) BIC, using $-\log f(y|\theta) + \frac{number\ params}{2} \log n$ [6].
(iv) MDL, using $-\log f(y|\theta) + \frac{continuous\ params}{2} \log n + \log \binom{n}{C}$.

5.1 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.

$\hat{k}$	1	2	3	4	5	Av. KL
	n=20					
MML	99	0	1	0	0	0.085
AIC	39	35	22	4	0	23.926
BIC	78	15	7	0	0	23.058
MDL	92	5	3	0	0	20.238
	n=40					
MML	98	2	0	0	0	0.033
AIC	30	20	31	14	5	9.089
BIC	87	10	3	0	0	7.487
MDL	98	2	0	0	0	0.424
	n=80					
MML	99	0	0	0	1	0.020
AIC	12	9	30	25	24	4.446
BIC	95	4	1	0	0	0.483
MDL	99	1	0	0	0	0.265
	n=160					
MML	99	1	0	0	0	0.007
AIC	6	9	23	31	31	3.961
BIC	99	1	0	0	0	0.088
MDL	100	0	0	0	0	0.007

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

$\hat{k}$	1	2	3	4	5	Av. KL
	n=20					
MML	69	28	3	0	0	0.324
AIC	15	47	30	8	0	24.172
BIC	48	38	9	5	0	23.510
MDL	70	21	6	3	0	23.061
	n=40					
MML	37	60	3	0	0	0.140
AIC	4	40	32	21	3	13.559
BIC	29	58	12	1	0	12.412
MDL	53	41	6	0	0	10.166
	n=80					
MML	11	81	6	1	1	0.088
AIC	0	17	27	30	26	7.246
BIC	16	76	7	0	1	0.816
MDL	34	63	3	0	0	0.770
	n=160					
MML	0	98	2	0	0	0.025
AIC	0	23	32	26	19	2.777
BIC	1	97	2	0	0	0.108
MDL	2	98	0	0	0	0.027

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

$\hat{k}$	1	2	3	4	5	6	Av. KL
	n=20						
MML	31	65	4	0	0	0	0.320
AIC	3	49	43	4	1	0	17.441
BIC	15	61	22	2	0	0	16.884
MDL	34	52	13	1	0	0	16.034
	n=40						
MML	3	85	12	0	0	0	0.191
AIC	0	28	41	26	4	1	10.379
BIC	3	79	16	1	1	0	9.337
MDL	10	78	10	1	1	0	9.255

$\hat{k}$	1	2	3	4	5	6	Av. KL
	n=80						
MML	0	50	50	0	0	0	0.106
AIC	0	4	34	35	23	4	6.089
BIC	0	61	36	3	0	0	2.786
MDL	0	77	22	1	0	0	1.358
	n=160						
MML	0	8	92	0	0	0	0.044
AIC	0	0	32	28	21	19	2.729
BIC	0	21	79	0	0	0	1.416
MDL	0	46	54	0	0	0	1.316

Table 2. True no. of segments = 3

5.2 Results

In Tables 1(a), 1(b) and 2, we give the results when we presented simulated data to the criteria given in Section 5. The data used in the simulations was generated according to the following distributions:

Table 1(a) — One segment with distribution $N(\mu = 0, \sigma^2 = 1)$,

Table 1(b) — Two segments with the first half distributed as $N(\mu = 0, \sigma^2 = 1)$ and the second half distributed as $N(\mu = 1, \sigma^2 = 1)$, and

Table 2 — Three segments with the first third distributed as $N(\mu = 0, \sigma^2 = 1)$,

the middle third distributed as $N(\mu = 1, \sigma^2 = 1)$ and the last third distributed as $N(\mu = 2, \sigma^2 = 1)$.
In each simulation, we generated n points from the appropriate distribution. We applied the search method described in Section 5.1. We applied the criteria from Section 5 and listed the number of times the criteria estimated each value of k from 100 simulations. Tables 1(a), 1(b) and 2 also give the average Kullback-Liebler distance (Av. KL) between the predicted distribution, and the underlying distribution[5].

6 Time Series Applications

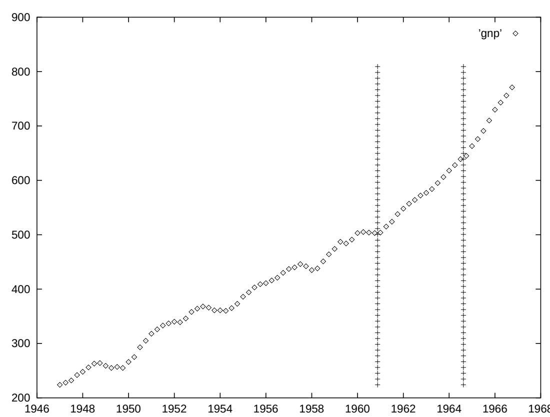

Fig. 2. The US GNP 1947 – 1966

We may model time series of the form: $z_{t+1} = z_t + c_j + \epsilon(0, \sigma_j^2)$ by setting $y_t = z_{t+1} - z_t$. This may be a reasonable method for segmenting data from examples such as: (i) economic time series, (ii) electrocardiogram measurements and (iii) eye movement measurements from a sleeping person.

We segmented the quarterly gross national product (GNP) for the United States from 1947 – 1966 [14]. Figure 2[6] shows the preferred MML segmentation for this data. The BIC and MDL criteria also preferred this segmentation, while the AIC criterion preferred a segmentation with 7 segments.

[5] The Kullback-Liebler distance (given for example in [15, Chp. 9]) between a true distribution $N(\mu_t,\ \sigma_t^2)$ and a fitted distribution $N(\mu_f,\ \sigma_f^2)$ is

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

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

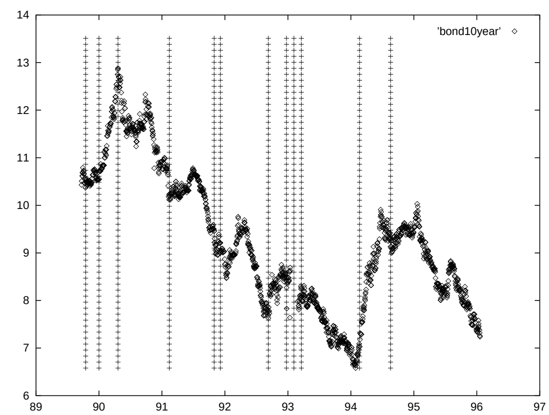

Fig. 3. The Canadian 10 year bond yield 1989 – 1996 with 12 cut points

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 MML criterion found evidence for there being at least 8 cut points 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 3) was 5282.8 nits.

7 Conclusion

We have derived a message length criterion for the segmentation of univariate data with Gaussian noise. We tested the criterion and found that it outperformed other criteria (AIC, BIC, MDL) in determining the number of regions in the simulations conducted here. Of the methods considered in this paper, the average Kullback-Liebler distance between the fitted distribution and true distribution was far smaller for the MML method. The method was successfully applied to two financial time series problems; the method scaled up reasonably to handle a data set with 1514 data points.

Acknowledgments

We would like to thank Catherine Forbes, David Albrecht and Wray Buntine for valuable discussions, and the anonymous referees for valuable critical comments. Jon Oliver acknowledges research support by Australian Research Council (ARC) Postdoctoral Research Fellowship F39340111.

References

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proc. of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.
2. 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, 1996.
3. J.H. Conway and N.J.A Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, London, 1988.
4. 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.
5. 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.*
6. Mengxiang Li. Minimum description length based 2-D shape description. In *IEEE 4th Int. Conf. on Computer Vision*, pages 512–517, May 1992.
7. 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.
8. J.J. Oliver and D.J. Hand. Introduction to minimum encoding inference. Technical report TR 4-94, Dept. of Statistics, Open University, Walton Hall, Milton Keynes, MK7 6AA, UK, 1994. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.
9. J.J. Oliver, Baxter R.A., and Wallace C.S. Unsupervised Learning using MML. In *Machine Learning: Proc. of the Thirteenth International Conference (ICML 96)*, pages 364–372. Morgan Kaufmann Publishers, San Francisco, CA, 1996. Available on the WWW from `http://www.cs.monash.edu.au/~jono`.
10. B. Pfahringer. Compression-based discretization of continuous attributes. In *Machine Learning: Proc. of the Twelfth International Workshop*, pages 456–463, 1995.
11. J.R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence*, 4:77–90, 1996.
12. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
13. G. Schwarz. Estimating dimension of a model. *Ann. Stat.*, 6:461–464, 1978.
14. 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.
15. C.W. Therrien. *Decision, estimation, and classification : an introduction to pattern recognition and related topics.* Wiley, New York, 1989.
16. H. Tong. *Non-linear time series : a dynamical system approach.* Clarendon Press, Oxford, 1990.
17. C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11:185–194, 1968.
18. 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.

This article was processed using the LaTeX macro package with LLNCS style