Minimum Message Length Grouping of Ordered Data

Leigh J. Fitzgibbon, Lloyd Allison, and David L. Dowe

School of Computer Science and Software Engineering
Monash University, Clayton, VIC 3168 Australia
{leighf,lloyd,dld}@csse.monash.edu.au

Abstract. Explicit segmentation is the partitioning of data into homogeneous regions by specifying cut-points. W. D. Fisher (1958) gave an early example of explicit segmentation based on the minimisation of squared error. Fisher called this *the grouping problem* and came up with a polynomial time Dynamic Programming Algorithm (DPA). Oliver, Baxter and colleagues (1996,1997,1998) have applied the information-theoretic Minimum Message Length (MML) principle to explicit segmentation. They have derived formulas for specifying cut-points imprecisely and have empirically shown their criterion to be superior to other segmentation methods (AIC, MDL and BIC). We use a simple MML criterion and Fisher's DPA to perform numerical Bayesian (summing and) integration (using message lengths) over the cut-point location parameters. This gives an estimate of the number of segments, which we then use to estimate the cut-point positions and segment parameters by minimising the MML criterion. This is shown to have lower Kullback-Leibler distances on generated data.

1 Introduction

Grouping is defined as the partitioning, or explicit segmentation, of a set of data into homogeneous groups that can be explained by some stochastic model [8]. Constraints can be imposed to allow only contiguous partitions over some variable or on data-sets that are ordered a priori. For example, time series segmentation consists of finding homogeneous segments that are contiguous in time.

Grouping theory has applications in inference and statistical description problems and there are many practical applications. For example, we wish to infer when and how many changes in a patient's condition have occurred based on some medical data. A second example is that we may wish to describe Central Processor Unit (CPU) usage in terms of segments to allow automatic or manager-based decisions to be made.

In this paper, we describe a Minimum Message Length (MML) [18, 22, 19] approach to explicit segmentation for data-sets that are ordered a priori. Fisher's original Maximum Likelihood solution to this problem was based on the minimisation of squared error. The problem with Maximum Likelihood approaches is that they have no stopping criterion, which means that unless the number of

H. Arimura, S. Jain and A. Sharma (Eds.): ALT 2000, LNAI 1968, pp. 56-70, 2000.

groups is known a priori, the optimal grouping would consist of one datum per group. Maximum Likelihood estimates for the cut-point positions are also known to be inaccurate [11] and have a tendency to place cut-points in close proximity of each other. MML inference overcomes both these problems by encoding the model and the data as a two-part message.

The MML solution we describe is based on Fisher's polynomial time Dynamic Programming Algorithm (DPA), which has several advantages over commonly used graph search algorithms. It is able to handle adjacent dependencies, where the cost of segment i is dependent on the model for segment $i-1$. The algorithm is exhaustive and can be made to consider all possible segmentations, allowing for numerical (summing and) integration. Computing the optimal segmentation of data into G groups results in the solution of all optimal partitions for $1..G$ over $1..K$, where K is the number of elements in the data-set.

Oliver, Baxter, Wallace and Forbes [3, 11, 10] have implemented and tested a MML based solution to the segmentation of time series data and compared it with some other techniques including Bayes Factors [9], AIC [1], BIC [15], and MDL [12]. In their work, they specify the cut-point to a precision that the data warrants. This creates dependencies between adjacent segments and without knowledge of Fisher's DPA they have used heuristic search strategies. They have empirically shown their criterion to be superior to AIC, BIC and MDL over the data-sets tested. However, the testing was only performed on data with fixed parameter values and equally spaced cut-points.

We use a simple MML criterion and Fisher's DPA to perform Bayesian (summing and) integration (using message lengths) over the cut-point parameter(s). This gives an estimate of the number of segments, which we then use to estimate the cut-point positions and segment parameters by minimising the MML criterion. This unorthodox[1] coding scheme has the advantage that because we do not state the cut-point positions, we do not need to worry about the precision to which they are stated and therefore reduce the number of assumptions and approximations involved. We compare our criterion with Oliver and Baxter's [11] MML, MDL and BIC criteria over a number of data-sets with and without randomly placed cut-points and parameters.

This paper is structured as follows. Section 2 contains background information on Fisher's grouping problem and his algorithm. It also contains an overview of the MML segmentation work by Oliver, Baxter and others [3, 11, 10] and an introduction to Minimum Message Length inference. Section 3 contains a re-statement of the segmentation problem using our terminology. In Section 4, we describe the message length formula that we use to segment the data and the approximate Bayesian integration technique we use to remove the cut-point parameter. In Section 5, we perform some experiments and compare with the previous work of Oliver, Baxter and others [10]. The concluding Sections, 6 and 7, summarize the results and suggest future work.

[1] Unorthodox in terms of the Minimum Message Length framework [18, 22, 19], where parameters that are to be estimated should be stated in the first part of the message.

2 Background

2.1 The Grouping Problem

An ordered set of K numbers $\{a_i : i = 0..K-1\}$ can be partitioned into G contiguous groups in $\binom{K-1}{G-1}$ ways. We only consider contiguous partitions since we assume that the data has been ordered a priori[2]. For a given G, Fisher's solution to the grouping problem was to search for the contiguous partition determined by $G-1$ cut-points that minimised the distance, D:

$$D = \sum_{i=0}^{K-1} (a_i - \overline{a}_i)^2 \tag{1}$$

where $\overline{a}_i$ represents the arithmetic mean of the a's assigned to the group in which i is assigned. For a given G, the partition which minimises D is called an optimal or least squares partition. Whilst Fisher was concerned with grouping normally distributed data (fitting piecewise constants), his techniques, and the techniques derived in this paper can be applied to other models.

The exhaustive search algorithm used to find the optimal partition is based on the following "Sub-optimisation Lemma"[8, page 795]:

Lemma 1. *If $A_1 : A_2$ denotes a partition of set A into two disjoint subsets A_1 and A_2, if P_1* denotes a least squares partition of A_1 into G_1 subsets and if P_2* denotes a least squares partition of A_2 into G_2 subsets; then, of the class of sub-partitions of $A_1 : A_2$ employing G_1 subsets over A_1 and G_2 subsets over A_2 a least squares sub-partition is $P_1* : P_2*$.*

This lemma is possible due to the additive nature of the distance measure. The algorithm based on this lemma is an example of a Dynamic Programming Algorithm (DPA) and is computable in polynomial time. The DPA is a general class of algorithm that is used in optimisation problems where the solution is the sum of sub-solutions. Fisher's algorithm can easily be expressed in pseudo-code. In Figure 1 the pseudo-code for a function $D(G)$ which returns the distance, D, for a number of groups, G, up to an upper bound G_{max} is shown.

The time complexity of Fisher's DPA is:

$$\forall_{k=1..G_{max}-1} \forall_{i=k..K-1} min_{j=k}^{i} D[k-1, j-1] + sumsqr(j,i) = O(G_{max} \cdot K^2) \tag{2}$$

In practice, $G_{max} \ll K$.

2.2 The Problem with the Maximum Likelihood Partition

How many segments? Given some data, where G is unknown, a practitioner must view a range of least square partition solutions and then select one. For easy

[2] This is what W. D. Fisher called the *restricted problem.*

```
Lookup functions:
$sum(i,j) = sum[j+1] - sum[i]$
$sumsqr(i,j) = sumsqr[j+1] - sum[i]$
$D(i,j) = sumsqr(i,j) - \frac{\mathrm{sum}(i,j)^2}{j-i+1}$
$D(G) = D[G-1, K-1]$
Boundary conditions:
$sum[0] := 0$
$sumsqr[0] := 0$
Initial Step:
$sum[i] := sum[i-1] + a_{i-1}, \forall_{i=1..K}$
$sumsqr[i] := sumsqr[i-1] + a_{i-1}^2, \forall_{i=1..K}$
$D[0,i] := D(0,i), \forall_{i=0..K-1}$
General Step:
$D[k,i] := min_{j=k}^{i} D[k-1, j-1] + sumsqr(j,i),$
$\forall_{k=1..G_{max}-1} \forall_{i=k..K-1}$
```

Fig. 1. A Dynamic Programming Algorithm based on Fisher's Sub-optimisation Lemma.

data this may be satisfactory. However, for difficult data a human cannot detect subtle differences between the solutions. Consider the least square partitions for $G = \{2,3,4,5\}$ of some generated data in Figures 3 to 6. From inspection of these four hypotheses, it is difficult to determine the true number of segments.

Poor parameter estimates Even when we know the number of segments in a data-set, the least squares partition may give poor estimates for the cut-point positions, and segment parameters. Oliver and Forbes [11] found that the Maximum Likelihood estimates for the cut-point position are unreliable. In their experiments the Maximum Likelihood technique that was given the correct number of segments had, on average, a higher Kullback-Leibler distance than a MML based technique that did not know the correct number of segments. An example of this can be seen in the least squares partitions in Figures 5 and 6. The least squares and MDL methods tend to place cut-points in close proximity of each other.

2.3 The Minimum Message Length Principle

The MML principle [18, 22, 19] is based on compact coding theory. It provides a criterion for comparing competing hypotheses (models) by encoding both the hypothesis and the data in a two-part message. For a hypothesis, H, and data, D, Bayes's theorem gives the following relationship between the probabilities:

$$Pr(H\&D) = Pr(H) \cdot Pr(D|H) = Pr(D) \cdot Pr(H|D), \tag{3}$$

which can be rearranged as:

$$Pr(H|D) = \frac{Pr(H) \cdot Pr(D|H)}{Pr(D)} \tag{4}$$

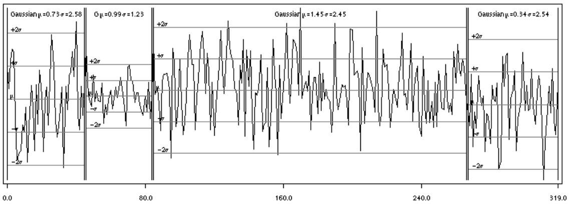

Fig. 2. Some generated data.

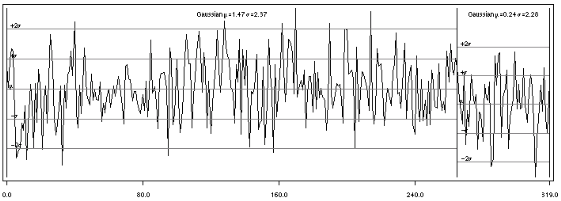

Fig. 3. Two segment least squares partition of Fig. 2.

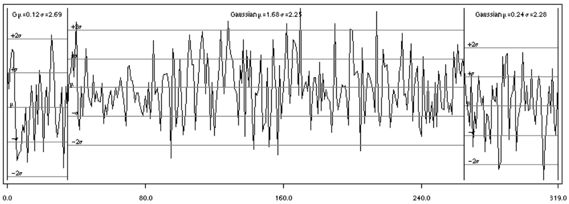

Fig. 4. Three segment least squares partition of Fig. 2.

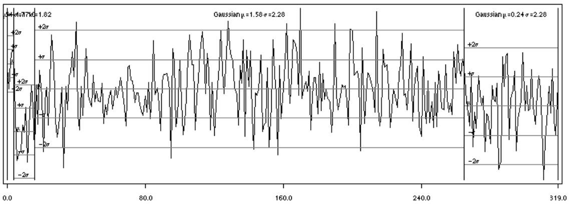

Fig. 5. Four segment least squares partition of Fig. 2.

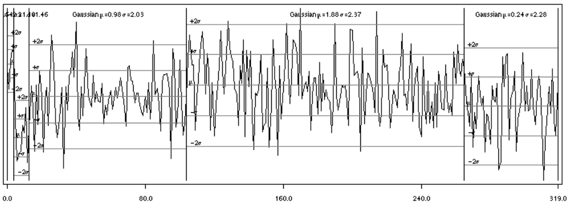

Fig. 6. Five segment least squares partition of Fig. 2.

After observing some data D, it follows that $Pr(H|D)$ is maximised when $Pr(H) \cdot Pr(D|H)$ is maximised. We know from coding theory that an event with probability P can be transmitted using an optimal code in a message of $-\log_2(P)$ bits[3] in length. Therefore the length of a two-part message (MessLen) conveying the parameter estimates (based on some prior) and the data encoded based on these estimates can be calculated as:

$$MessLen(H\&D) = -\log_2(Pr(H)) - \log_2(Pr(D|H)) \text{ bits} \tag{5}$$

The receiver of such a hypothetical message must be able to decode the data without using any other knowledge. Minimising $MessLen(H\&D)$ is equivalent to maximising $Pr(H|D)$, the latter being a *probability* and *not* a density [20, section 2] [21, section 2] [5]. The model with the shortest message length is considered to give the best explanation of the data. This interpretation of inductive inference problems as coding problems has many practical and theoretical advantages over dealing with probabilities directly. A survey of MML theory and its many successful applications is given by Wallace and Dowe [19].

2.4 MML Precision of Cut-point Specification

We can encode the cut-point positions in $\log\binom{K-1}{G-1}$ nits. However, using this coding scheme can be inefficient for small sample sizes and noisy data. Consider two segments whose boundaries are not well-defined: the posterior distribution will not have a well defined mode, but there may be a region around the boundary with high probability. The MML principle states that we should use this region to encode the data - we should only state the cut-point to an accuracy that the data warrants, for otherwise we risk under-fitting.

Oliver, Baxter and others [3, 11, 10] studied the problem of specifying the cut-point imprecisely. They derived equations to calculate the optimal precision with which to specify the cut-point. Where the boundary between two segments is not well-defined, it is cheaper to use less precision for the cut-point specification. This reduces the length of the first part of the message but may increase the length of the second part. Where the boundary is well-defined, it pays to use a higher precision to save in the second part of the message. Empirical results [3, 11, 10] have shown that specifying cut-points imprecisely gives better estimates of the number of segments and lower Kullback-Leibler distances. Similar success with MML imprecise cut-point specification has been found by Viswanathan, Wallace, Dowe and Korb [17] for binary sequences.

3 Problem Re-Statement

We consider a process which generates an ordered data-set. The process can be approximated by, or is considered to consist of, an exhaustive concatenation of contiguous sub-sets that were generated by sub-processes. We consider a

[3] In the next sections of the paper we use the natural logarithm and the unit is nits.

sub-process to be homogeneous and the data generated by a process to consist entirely of one or more homogeneous sequences.

Let y be a univariate ordered data-set of K numbers generated by some process:

$$y = (y_0, y_1, ..., y_{K-1}) \tag{6}$$

which consists of G contiguous, exhaustive and mutually exclusive sub-sets:

$$s = (s_0, s_1, ..., s_{G-1}), \tag{7}$$

where the members of each s_i were generated by sub-process i, which can be modelled with parameters θ_i:

$$\theta = (\theta_0, \theta_1, ..., \theta_{G-1}), \tag{8}$$

and likelihood:

$$f(y \in s_i | \theta_i) \tag{9}$$

In some cases, the number of distinct sub-processes may be less than G. This is most likely to occur in processes that have discrete states. For example, a process that alternates between two discrete states would be better modelled as coming from two, rather than G, sub-processes since parameters would be estimated over more data. This is a common approach with implicit segmentation, where segments are modelled implicitly by a Markov Model [16, 7]. However, the use of G sub-processes results in a more tractable problem and is what is generally used for explicit segmentation. Moreover, in some cases we may wish to model data which can be considered as coming from a drifting process rather than a process with distinct states. In these cases, segmentation can be used to identify approximately stationary regions and is best modelled as coming from G distinct sub-processes.

The inference problem is to estimate some or all of : G, s, θ and $f(y \in s_i|\theta_i)$.

4 Calculating the Message Length with Gaussian Segments

In this section we describe the message length formula used to calculate the expected length of a message which transmits the model and the data. Assume that the size, K, of the data-set is known and given. In order for a hypothetical receiver to decode the message and retrieve the original data, we must encode the following: G, the number of segments; the cut-point positions, $c|s$; the parameter estimates, θ_i, for each segment s_i; and finally the data for each segment using the parameter estimates stated. We specify G using the universal $\log^*$ code [13, 2], although we re-normalise the probabilities because we know that $G \leq K$. This simplifies the problem to the specification of:

– the cut-point positions $c|s$,
– the parameter estimates θ_i and data for each segment.

From Wallace and Freeman [22], the formula for calculating the length of a message where the model consists of several continuous parameters $\theta = (\theta_1, \ldots, \theta_n)$ is:

$$MessLen(H\&D) = -\log\left(\frac{h(\theta)f(y|\theta)}{\sqrt{F(\theta)}}\right) + \frac{n}{2}(1 + \log\kappa_n) \text{ nits} \qquad (10)$$

where $h(\theta)$ is a prior distribution over the n parameter values, $f(y|\theta)$ is the likelihood function for the model, $F(\theta)$ is the determinant of the Fisher Information matrix and κ_n is a lattice constant which represents the saving over the quantised n-dimensional space.

In this paper we consider Gaussian segments with two continuous parameters μ and σ: $y \in s_j \sim N[\mu_j, \sigma_j]$ so, $\theta_j = (\mu_j, \sigma_j)$. The lattice constant $\kappa_2 = \frac{5}{36\sqrt{3}}$ [4], the Fisher Information, $F(\theta)$, for the Normal distribution [10] is:

$$F(\mu, \sigma) = \frac{2n^2}{\sigma^4} \qquad (11)$$

and the negative log-likelihood is:

$$-\log f(y|\mu, \sigma) = \frac{n}{2}\log 2\pi + n\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (12)$$

The prior distribution we use is non-informative based on the population variance, $\sigma^2_{pop} = \frac{1}{K-1}\sum_{i=0}^{K-1}(y_i - \mu_{pop})^2$ where $\mu_{pop} = \frac{1}{K}\sum_{i=0}^{K-1} y_i$:

$$\forall_j \; h(\mu_j, \sigma_j) = \frac{1}{2\sigma^2_{pop}} \qquad (13)$$

This is the prior used by Oliver, Baxter and others [11, section 3.1.3] [3, 10], although the prior $\forall_j \; h(\mu_j, \sigma_j) = \frac{1}{4\sigma^2_{pop}}$ from [18, section 4.2] or other priors could also be considered. We use this prior, from Equation 13, to allow for a fair comparison with their criterion [3, 11, 10].

We use Equation 10 to send the parameters $\theta_j = (\mu_j, \sigma_j)$ and data for each segment. To encode the cut-point positions we use a simple coding scheme assuming that each combination is equally likely:

$$MessLen(c|K, G) = \log\binom{K-1}{G-1} \text{ nits} \qquad (14)$$

Based on Equation 10, the expected total length of the message is:

$$\begin{aligned} MessLen(H\&D) &= \log^*(G) + MessLen(c|K, G) \qquad (15) \\ &+ \sum_{j=1}^{G}\left(-\log\left(\frac{h(\theta_j)f(y \in s_j|\theta_j)}{\sqrt{F(\theta_j)}}\right) + \frac{n}{2}(1 + \log\kappa_n)\right) \text{ nits} \end{aligned}$$

If we were to optimise the values of G, s and θ to minimise Equation 15, we would under-estimate G since c is being stated to maximum precision (see Section 2.4). We avoid this problem by summing the probabilities of the various MML estimates of $\theta_j = (\mu_j, \sigma_j)_{j=0,..,G-1}$ over all possible sub-partitions:

$$Prob'(G) = \sum_{i=1}^{\binom{K-1}{G-1}} e^{-MessLen(H\&D)_i} \tag{16}$$

where $MessLen(H\&D)_i$ is the message length associated with the ith sub-partition from the $\binom{K-1}{G-1}$ possible sub-partitions and the values of the $\hat{\theta}_j$ associated with each such ith sub-partition. $Prob'$ gives unnormalised probabilities for the number of segments. The 'probabilities' are unnormalised because, for each ith sub-partition, the 'probabilities' consider only that part of the posterior density of the θ_j contained in the MML coding block[4].

We optimise Equation 16 to estimate G. This can be implemented by modifying Fisher's DPA given in Figure 1 by replacing the distance function with Equation 10 and changing the general step to sum over all sub-partitions:

$$\begin{aligned} D[k,i] :=& LOGPLUS(D[k-1, j-1], sumsqr(j,i)) \\ & \forall_{k=1..G_{max}-1} \forall_{i=k..K-1} \forall_{j=k..i} \end{aligned} \tag{17}$$

where the $LOGPLUS$ function is used to sum the log-probabilities:

$$LOGPLUS(x,y) = -\log_e(e^{-x} + e^{-y}) \tag{18}$$

Using Equation 16 to estimate G we then optimise Equation 15 to estimate the remaining parameters.

5 Experimental Evaluation

5.1 Generated Data

We now use Fisher's DPA to infer the number of segments G, the cut-point positions $c|s$ and segment parameters θ_i of some generated Gaussian data. The criteria to be compared are:

- MML-I, Equations 15 and 16 from the previous section.
- MMLOB, MML Equation (6) from the paper Oliver and Baxter [10].
- BIC, using $-\log f(x|\theta) + \frac{numberparams}{2} \log K$.
- MDL, using $-\log f(x|\theta) + \frac{continuousparams}{2} \log n + \log \binom{K}{G}$.

[4] However, normalising these 'probabilities' will give a reasonable approximation [5, sections 4 and 4.1] [19, sections 2 and 8] to the marginal/posterior probability of G which would be obtained by integrating out over all the $\theta_j = (\mu_j, \sigma_j)$.

The BIC and MDL criteria[5] were included since these were investigated and compared by Oliver and Baxter [10, page 8], but not over the range of data that we consider. AIC was omitted due to its poor performance in previous papers [3, 11, 10]. We expect our criterion, MML-I, to perform better where the data is noisy, the sample size is small or where the approximations break down in MMLOB.

We have generated three different data-sets S_0, S_1 and S_2:

- S_0 has fixed μ's and σ's and evenly-spaced cut-points; similar to Oliver and Baxter [10].
- S_1 has fixed μ's and σ's and (uniformly) randomly chosen cut-points (minimum segment size of 3).
- S_2 has random μ's and σ's drawn uniformly from [0..1], and (uniformly) randomly chosen cut-points (minimum segment size of 3).

For each data-set, 100 samples were generated of sizes 20, 40, 80, 160 and 320 and with each of 1..7 segments. For S_0 and S_1, the variance of each segment is 1.0, and the means of the segments are monotonically increasing by 1.0.

5.2 Experimental Results

We have collated the data collected during the experiments to report: a count of the number of times the correct number of cut-points were inferred (score test); the average number of cut-points inferred; and the Kullback-Leibler (KL) distance between the true and inferred distribution. The KL distance gives an indication of how well the parameters for each segment are being estimated. This will be affected by the inferred number of cut-points and their placement.

MDL and BIC were generally out-performed by the two MML methods (MML-I and MMLOB) in all measures. The interesting comparison is between MML-I and MMLOB.

Not all of the results could be included due to space limitations. The KL distance and average number of cut-points for S_0 and S_1 were omitted. For these two data-sets, the average number of inferred cut-points was slightly better for MML-I, and the KL distances for MML-I and MMLOB were both very similar.

The score test results have been included for all data-sets and can be seen in Tables 1 to 2. Each table shows the number of times the correct number of cuts k was inferred from the 100 trials for each of the sample sizes under investigation (20,40,80,160 and 320). MML-I is more accurate than the other criteria for both S_0 and S_1 on the score test. The strange exception is for S_2, where MMLOB is not only more accurate than the other criteria, but has improved a seemingly disproportionate amount over its results for S_0 and S_1.

Table 3 shows the average number of inferred cuts for data-set S_2. None of the criteria appear to be excessively over-fitting.

[5] We also note that MDL has been refined [14] since the 1978 MDL paper [12]. For a general comparison between MDL and MML, see, e.g., [14, 19, 20] and other articles in that special issue of the *Computer Journal.*

Table 4 shows the average Kullback-Leibler (KL) distances and standard deviations for data-set S_2. The KL distance means and standard deviations for MML-I are consistent for all sample sizes and are overall best, performing exceptionally well on sample sizes $K \leq 40$. MMLOB, MDL and BIC appear to break down for small samples in terms of both the mean and standard deviation.

MML-I has consistently low KL distances over all data-sets and is generally able to more accurately infer the number of cut-points for S_0 and S_1 than the other criteria. MMLOB is more accurate at inferring the number of cuts for data-set S_2 but has substantially higher KL distances than MML-I, but slightly better KL distances than BIC and MDL.

5.3 Application to Lake Michigan-Huron Data

We have used the MML-I criterion developed in this paper to segment the lake Michigan-Huron data that was posed as a problem in W. D. Fisher's original 1958 paper [8]. The DPA using our criterion was implemented in Java 2 (JIT) and was able to consider the over 10^{12} possible segmentations (for $G \leq 10$) of the lake data, with $K = 96$ in 2.1 seconds on a Pentium running at 200 mega-hertz. It inferred that there are five segments; $G = 5$. A graph of the segmentation can be seen in Figure 7. In Figure 8 we have segmented the lake data up to the year 1999. We can see that the segmentation identified in Figure 7 has been naturally extended in Figure 8.

Fisher's original least squares program was written for the "Illiac" digital computer at the University of Illinois and could handle data-sets with $K \leq 200$ and $G \leq 10$ with running time up to approximately 14 minutes.

6 Conclusion

We have applied numerical Bayesian (summing and) integration for cut-point parameters in the grouping or segmentation problem. Using W. D. Fisher's polynomial time DPA, we were able to perform approximations to numerical Bayesian integration using a Minimum Message Length criterion (MML-I) to estimate the number of segments. Having done that, we then minimize the MML-I criterion (Equation 15) to estimate the segment boundaries and within-segment parameter values. This technique, MML-I, was compared with three other criteria: MMLOB [11], MDL and BIC. The comparison was based on generated data with fixed and random parameter values. Using the Fisher DPA, we were able to experiment over a larger range of data than previous work [3,11,10]. The MMLOB and MML-I criteria performed well and were shown to be superior to MDL and BIC. The MML-I criterion, using Bayesian integration, was shown to have overall lower Kullback-Leibler distances and was generally better at inferring the number of cut-points than the other criteria.

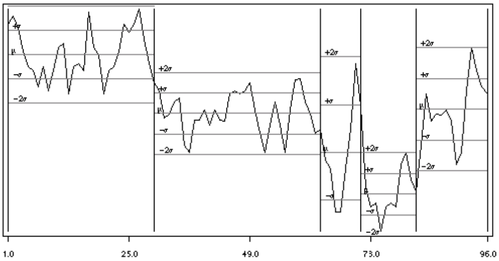

Fig. 7. Lake Michigan-Huron monthly mean water levels from 1860 to 1955 segmented by MML-I. This is the data that W. D. Fisher originally considered in 1958.

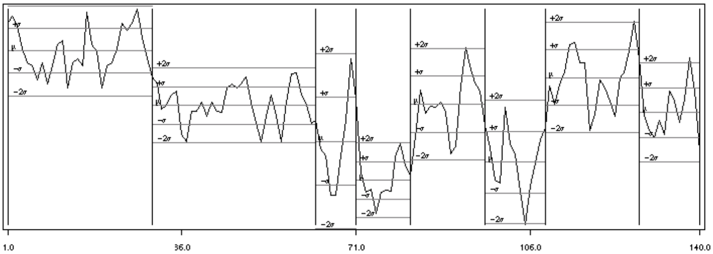

Fig. 8. Lake Michigan-Huron monthly mean water levels from 1860 to 1999 segmented by MML-I.

Table 1. Positive inference counts for data-set S_0.

$\hat{k}$	Criterion	20	40	80	160	320	Total
0	MML-I	86	93	93	100	95	467
	MMLOB	93	94	93	100	84	464
	MDL	89	96	99	100	99	483
	BIC	76	88	95	98	96	453
1	MML-I	43	69	76	83	89	360
	MMLOB	28	57	86	89	77	337
	MDL	24	35	62	96	98	315
	BIC	42	55	83	89	96	365
2	MML-I	3	21	63	74	91	252
	MMLOB	6	12	46	84	81	229
	MDL	4	10	13	52	98	177
	BIC	11	23	35	68	92	229
3	MML-I	0	3	17	51	79	150
	MMLOB	1	3	10	61	68	143
	MDL	2	5	5	14	76	102
	BIC	2	9	9	34	88	142
4	MML-I	0	0	6	44	77	127
	MMLOB	0	0	2	22	65	89
	MDL	0	1	1	2	45	49
	BIC	0	4	7	11	58	80
5	MML-I	0	0	0	19	66	85
	MMLOB	0	0	0	7	64	71
	MDL	0	0	0	2	9	11
	BIC	0	1	0	9	21	31
6	MML-I	0	0	0	5	49	54
	MMLOB	0	0	0	0	56	56
	MDL	0	0	0	0	3	3
	BIC	0	0	0	1	8	9

Table 2. Positive inference counts for data-sets S_1 and S_2 respectively.

$\hat{k}$	Criterion	20	40	80	160	320	Total
1	MML-I	32	45	64	74	77	292
	MMLOB	22	40	62	72	78	274
	MDL	19	23	38	67	79	226
	BIC	35	38	57	78	81	289
2	MML-I	5	18	34	50	65	172
	MMLOB	5	6	27	43	58	139
	MDL	4	2	12	27	48	93
	BIC	14	9	23	37	54	137
3	MML-I	0	1	8	29	48	86
	MMLOB	0	1	9	16	50	76
	MDL	2	4	3	4	20	33
	BIC	3	8	13	14	32	70
4	MML-I	0	0	4	14	32	50
	MMLOB	0	0	3	12	30	45
	MDL	0	0	0	1	2	3
	BIC	0	2	2	6	12	22
5	MML-I	0	0	0	5	17	22
	MMLOB	0	0	0	1	24	25
	MDL	0	1	1	0	2	4
	BIC	0	1	1	1	9	12
6	MML-I	0	0	0	0	7	7
	MMLOB	0	0	0	0	9	9
	MDL	0	1	0	0	0	1
	BIC	0	1	0	0	2	3

$\hat{k}$	Criterion	20	40	80	160	320	Total
1	MML-I	37	46	55	60	79	277
	MMLOB	44	56	65	68	82	315
	MDL	42	49	54	68	82	295
	BIC	49	51	59	69	85	313
2	MML-I	11	26	27	50	43	157
	MMLOB	16	33	35	57	57	198
	MDL	19	27	28	41	45	160
	BIC	30	37	37	52	53	209
3	MML-I	0	9	19	38	41	107
	MMLOB	2	12	30	37	51	132
	MDL	3	7	16	24	33	83
	BIC	5	11	27	30	40	113
4	MML-I	0	0	11	23	24	58
	MMLOB	0	5	10	24	27	66
	MDL	0	4	7	10	20	41
	BIC	0	6	14	15	25	60
5	MML-I	0	0	8	14	19	41
	MMLOB	0	1	9	13	28	51
	MDL	0	1	4	8	7	20
	BIC	0	1	4	12	12	29
6	MML-I	0	0	4	9	18	31
	MMLOB	0	0	3	9	20	32
	MDL	0	1	0	2	4	7
	BIC	0	1	1	5	9	16

Table 3. Average inferred number of cuts for data-set S_2.

$\hat{k}$	Criterion	20	40	80	160	320
0	MML-I	0.150 ± 0.39	0.100 ± 0.39	0.090 ± 0.38	0.000 ± 0.00	0.130 ± 0.77
	MMLOB	0.080 ± 0.31	0.100 ± 0.41	0.100 ± 0.41	0.000 ± 0.00	0.450 ± 1.50
	MDL	0.130 ± 0.39	0.040 ± 0.20	0.010 ± 0.10	0.000 ± 0.00	0.010 ± 0.10
	BIC	0.340 ± 0.67	0.210 ± 0.64	0.070 ± 0.33	0.030 ± 0.22	0.060 ± 0.34
1	MML-I	0.490 ± 0.61	0.640 ± 0.64	0.890 ± 0.82	0.960 ± 0.78	1.040 ± 0.85
	MMLOB	0.480 ± 0.54	0.660 ± 0.57	0.800 ± 0.65	0.880 ± 0.61	1.020 ± 0.67
	MDL	0.560 ± 0.62	0.570 ± 0.57	0.630 ± 0.60	0.720 ± 0.49	0.840 ± 0.39
	BIC	0.730 ± 0.66	0.830 ± 0.68	0.800 ± 0.64	0.820 ± 0.56	0.870 ± 0.37
2	MML-I	0.530 ± 0.69	0.980 ± 0.89	1.360 ± 1.04	1.640 ± 0.94	1.870 ± 1.28
	MMLOB	0.730 ± 0.76	1.100 ± 0.86	1.170 ± 0.79	1.580 ± 0.77	1.760 ± 0.91
	MDL	0.750 ± 0.80	1.010 ± 0.88	1.020 ± 0.82	1.220 ± 0.75	1.380 ± 0.71
	BIC	1.110 ± 0.82	1.270 ± 0.87	1.270 ± 0.87	1.440 ± 0.73	1.490 ± 0.72
3	MML-I	0.460 ± 0.64	1.060 ± 0.97	1.880 ± 1.26	2.510 ± 1.27	2.830 ± 1.43
	MMLOB	0.790 ± 0.87	1.260 ± 1.04	1.950 ± 1.10	2.170 ± 0.89	2.750 ± 0.99
	MDL	0.890 ± 0.91	1.100 ± 0.96	1.620 ± 1.06	1.780 ± 0.95	2.090 ± 0.84
	BIC	1.220 ± 0.91	1.440 ± 1.09	2.010 ± 1.11	2.000 ± 0.92	2.320 ± 0.85
4	MML-I	0.400 ± 0.60	1.010 ± 0.94	2.180 ± 1.27	3.050 ± 1.79	3.590 ± 1.54
	MMLOB	0.750 ± 0.86	1.360 ± 1.24	2.410 ± 1.16	2.750 ± 1.39	3.600 ± 1.38
	MDL	0.860 ± 0.96	1.130 ± 1.12	1.980 ± 1.08	2.080 ± 1.18	2.620 ± 1.03
	BIC	1.120 ± 0.97	1.540 ± 1.10	2.350 ± 1.03	2.370 ± 1.12	2.900 ± 1.08
5	MML-I	0.330 ± 0.62	1.080 ± 1.17	2.260 ± 1.46	3.500 ± 1.85	3.970 ± 1.62
	MMLOB	0.640 ± 0.92	1.690 ± 1.33	2.510 ± 1.49	3.210 ± 1.37	4.310 ± 1.53
	MDL	0.880 ± 1.02	1.510 ± 1.34	1.970 ± 1.37	2.490 ± 1.38	2.980 ± 1.14
	BIC	1.170 ± 1.02	1.910 ± 1.31	2.310 ± 1.33	2.870 ± 1.30	3.300 ± 1.14
6	MML-I	0.340 ± 0.61	1.200 ± 1.30	2.530 ± 1.69	3.660 ± 1.75	5.420 ± 1.96
	MMLOB	0.640 ± 0.86	1.810 ± 1.46	2.910 ± 1.56	3.610 ± 1.46	5.010 ± 1.56
	MDL	0.730 ± 0.90	1.550 ± 1.48	2.060 ± 1.26	2.820 ± 1.43	3.530 ± 1.23
	BIC	1.100 ± 0.96	1.930 ± 1.39	2.680 ± 1.28	3.280 ± 1.36	3.900 ± 1.21

Table 4. Kullback-Leibler distances for data-set S_2.

$\hat{k}$	Criterion	20	40	80	160	320
0	MML-I	0.218 ± 0.65	0.056 ± 0.13	0.023 ± 0.05	0.007 ± 0.01	0.013 ± 0.07
	MMLOB	0.422 ± 1.69	0.283 ± 2.08	0.034 ± 0.11	0.007 ± 0.01	0.049 ± 0.28
	MDL	0.716 ± 2.40	0.288 ± 2.10	0.016 ± 0.03	0.007 ± 0.01	0.007 ± 0.04
	BIC	1.159 ± 3.08	0.820 ± 3.24	0.132 ± 1.05	0.064 ± 0.52	0.046 ± 0.27
1	MML-I	0.588 ± 2.14	0.261 ± 0.29	0.154 ± 0.23	0.173 ± 0.55	0.072 ± 0.48
	MMLOB	4.650 ± 39.11	0.312 ± 0.90	0.389 ± 2.41	0.234 ± 1.56	0.070 ± 0.48
	MDL	5.633 ± 39.70	0.410 ± 1.36	0.538 ± 2.94	0.076 ± 0.24	0.137 ± 0.84
	BIC	5.841 ± 39.69	0.698 ± 2.04	0.725 ± 3.21	1.133 ± 9.52	0.136 ± 0.84
2	MML-I	0.542 ± 0.55	0.334 ± 0.30	0.244 ± 0.36	0.159 ± 0.30	0.227 ± 1.13
	MMLOB	0.835 ± 1.62	0.447 ± 1.22	0.248 ± 0.74	0.119 ± 0.21	0.145 ± 1.05
	MDL	1.590 ± 4.13	1.255 ± 7.05	0.260 ± 0.70	0.086 ± 0.14	0.035 ± 0.05
	BIC	1.625 ± 3.45	0.759 ± 1.70	1.022 ± 4.76	0.196 ± 0.59	0.045 ± 0.07
3	MML-I	0.620 ± 0.45	0.444 ± 0.40	0.266 ± 0.23	0.186 ± 0.22	0.116 ± 0.25
	MMLOB	1.181 ± 3.24	0.761 ± 2.61	0.322 ± 0.57	0.122 ± 0.22	0.097 ± 0.31
	MDL	1.323 ± 3.27	0.455 ± 0.61	0.470 ± 0.99	0.132 ± 0.27	0.085 ± 0.31
	BIC	1.650 ± 3.62	0.754 ± 1.18	0.909 ± 1.93	0.154 ± 0.32	0.169 ± 0.67
4	MML-I	0.670 ± 0.48	0.507 ± 0.48	0.361 ± 0.28	0.274 ± 0.43	0.176 ± 0.28
	MMLOB	5.499 ± 40.13	1.141 ± 4.70	0.454 ± 1.30	0.854 ± 6.52	0.542 ± 3.08
	MDL	6.013 ± 40.21	1.077 ± 4.64	0.518 ± 1.29	0.873 ± 6.51	0.279 ± 1.91
	BIC	3.710 ± 13.51	1.188 ± 3.30	0.760 ± 1.72	0.753 ± 4.09	1.255 ± 7.69
5	MML-I	0.671 ± 0.38	0.562 ± 0.33	0.441 ± 0.41	0.231 ± 0.20	0.202 ± 0.27
	MMLOB	3.826 ± 25.00	1.424 ± 6.27	0.572 ± 1.18	1.181 ± 9.36	0.133 ± 0.15
	MDL	2.096 ± 4.69	4.298 ± 25.49	0.755 ± 3.86	1.173 ± 9.36	0.118 ± 0.15
	BIC	2.476 ± 4.78	4.554 ± 25.49	0.803 ± 3.89	0.722 ± 3.49	0.240 ± 0.87
6	MML-I	0.722 ± 0.36	0.618 ± 0.48	0.386 ± 0.24	0.247 ± 0.19	0.299 ± 0.43
	MMLOB	5.688 ± 41.22	3.733 ± 24.12	0.674 ± 1.57	0.383 ± 1.03	0.259 ± 0.62
	MDL	5.756 ± 41.21	4.930 ± 28.90	0.816 ± 1.81	0.994 ± 4.87	0.169 ± 0.42
	BIC	4.375 ± 22.72	3.160 ± 10.83	1.206 ± 2.49	1.223 ± 4.92	0.294 ± 1.23

7 Further Work and Acknowledgments

We have not directly investigated how well the various criteria are placing the cut-points. The Kullback-Leibler distance gives an indirect measure since it is affected by the cut-point positions. We intend to perform a more explicit investigation into the placement of cut-points.

As well as the Gaussian distribution, MML formulas have been derived for discrete multi-state [17], Poisson, von Mises circular, and spherical Fisher distributions [21, 6]. Some of these distributions and other models will be incorporated in the future.

We thank Dean McKenzie for introducing us to the W. D. Fisher (1958) paper and Rohan Baxter and Jonathan Oliver for providing access to the C code used in Baxter, Oliver and Wallace [10].

References

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceeding 2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.
2. R. A. Baxter and J. J. Oliver. MDL and MML: Similarities and differences. Technical report TR 207, Dept. of Computer Science, Monash University, Clayton, Victoria 3168, Australia, 1994.

3. R. A. Baxter and J. J. Oliver. The kindest cut: minimum message length segmentation. In S. Arikawa and A. K. Sharma, editors, *Proc. 7th Int. Workshop on Algorithmic Learning Theory*, volume 1160 of *LCNS*, pages 83–90. Springer-Verlag Berlin, 1996.
4. J.H. Conway and N.J.A Sloane. *Sphere Packings, Lattices and Groups.* Springer-Verlag, London, 1988.
5. D. L. Dowe, R. A. Baxter, J. J. Oliver, and C. S. Wallace. Point estimation using the Kullback-Leibler loss function and MML. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD98)*, volume 1394 of *LNAI*, pages 87–95, 1998.
6. D. L. Dowe, J. J. Oliver, and C. S. Wallace. MML estimation of the parameters of the spherical Fisher distribution. In S. Arikawa and A. K. Sharma, editors, *Proc. 7th Int. Workshop on Algorithmic Learning Theory*, volume 1160 of *LCNS*, pages 213–227. Springer-Verlag Berlin, 1996.
7. T. Edgoose and L. Allison. MML markov classification of sequential data. *Statistics and Computing*, 9:269–278, 1999.
8. W. D. Fisher. On grouping for maximum homogeneity. *Jrnl. Am. Stat. Soc.*, 53:789–798, 1958.
9. R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
10. J. J. Oliver, R. A. Baxter, and C. S. Wallace. Minimum message length segmentation. In X. Wu, R. Kotagiri, and K. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining (PAKDD-98)*, pages 83–90. Springer, 1998.
11. J. J. Oliver and C. S. Forbes. Bayesian approaches to segmenting a simple time series. Technical Report 97/336, Dept. Computer Science, Monash University, Australia 3168, December 1997.
12. J. J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
13. J. J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2):416–431, 1983.
14. J. J. Rissanen. Hypothesis selection and testing by the MDL principle. *Computer Jrnl.*, 42(4):260–269, 1999.
15. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
16. S. Sclove. Time-series segmentation: A model and a method. *Information Sciences*, 29:7–25, 1983.
17. M. Viswanathan, C.S. Wallace, D.L. Dowe, and K. Korb. Finding cutpoints in noisy binary sequences - a revised empirical evaluation. In *12th Australian Joint Conference on Artificial Intelligence*, 1999. A sequel has been submitted to Machine Learning Journal.
18. C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Jrnl.*, 11(2):185–194, August 1968.
19. C. S. Wallace and D. L. Dowe. Minimum message length and Kolmogorov complexity. *Computer Jrnl.*, 42(4):270–283, 1999.
20. C. S. Wallace and D. L. Dowe. Rejoinder. *Computer Jrnl.*, 42(4):345–357, 1999.
21. C. S. Wallace and D. L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, 2000.
22. C. S. Wallace and P. R. Freeman. Estimation and inference by compact encoding (with discussion). *Journal of the Royal Statistical Society series B*, 49:240–265, 1987.