MDL BASED DIGITAL SIGNAL SEGMENTATION

Ciprian Doru Giurcăneanu, Ioan Tăbuş and Jorma Rissanen
Signal Processing Lab., Tampere University of Technology,
P.O.Box 553, Tampere, FINLAND
Tel: +358 3 3653911; fax: +358 3 3653857
e-mail: {cipriand,tabus}@cs.tut.fi,rissanen@almaden.ibm.com

ABSTRACT

The segmentation of a signal based on a piecewise polynomial model is reexamined here in light of the recent advances in applying the MDL principle to the exponential density family. The critical case of discriminating between short segments is handled very efficiently by using the exact MDL formula for small sample size, whereas MAP or AIC methods result in drastic over-segmentation. The simulation result for ECG segmentation shows that the piecewise linear approximation obtained with the proposed method preserves well the location of the QRS complex.

1 INTRODUCTION

The segmentation problem we address in this paper refers to finding the transition times when observing n samples $y_0, y_1, \ldots, y_{n-1}$ of a piecewise polynomial signal in additive gaussian noise. At unknown time instants, $T_1, \ldots, T_{r-1}$, some parameters in the model (polynomial coefficients or noise mean and variance) changes abruptly. The observed samples in each segment i are represented as

$$y_t = \underline{\beta}_i' \underline{x}_t + \varepsilon_t, \quad T_{i-1} \leq t < T_i, \quad T_0 = 1 \tag{1}$$

where $\underline{\beta}_i$ is the k_i-dimensional regressor vector, the regression vector is $\underline{x}_t = [1 \quad (t-T_{i-1})^1 \quad (t-T_{i-1})^{k_i-1}]'$ for $T_{i-1} \leq t < T_i$, r is the total number of segments, and $T_0 = 0$, $T_r = n$. The symbol $'$ denotes transposition and the observation vector is denoted $\underline{y}_0^n = [y_0 \; y_1 \; \ldots \; y_{n-1}]'$. The $n_i = T_i - T_{i-1}$ observations in the i-th segment can be written in vector form:

$$\underbrace{\begin{bmatrix} y_{T_{i-1}} \\ y_{T_{i-1}+1} \\ \vdots \\ y_{T_i-1} \end{bmatrix}}_{\underline{y}_{T_{i-1}}^{T_i}} = X_i \underbrace{\begin{bmatrix} \beta_{1,i} \\ \beta_{2,i} \\ \vdots \\ \beta_{k_i,i} \end{bmatrix}}_{\underline{\beta}_i} + \underbrace{\begin{bmatrix} \varepsilon_{T_{i-1}} \\ \varepsilon_{T_{i-1}+1} \\ \vdots \\ \varepsilon_{T_i-1} \end{bmatrix}}_{\underline{\varepsilon}_i}$$

where the polynomial on segment i has degree $k_i - 1$. The entries $\varepsilon_{T_{i-1}}, \varepsilon_{T_{i-1}+1}, \ldots, \varepsilon_{T_i-1}$ of the noise vector $\underline{\varepsilon}_i$ are i.i.d. samples from a zero mean gaussian source with variance $\tau_i = \sigma_i^2$ and the regression matrix is

$$X_i = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1^1 & \ldots & 1^{k_i-1} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & (n_i-1)^1 & \ldots & (n_i-1)^{k_i-1} \end{bmatrix}$$

Given $y_0, y_1, \ldots, y_{n-1}$, the objective is to estimate the number of segments, their boundaries, the degrees and coefficients of polynomials and the noise variance for each segment.

The hypothesis of zero mean Gaussian noise implies that the measurements for each segment i are distributed according to the probability density function (pdf) $f_{T_{i-1},T_i,k_i}(\underline{y}_{T_{i-1}}^{T_i} | \underline{\beta}_i, \tau_i)$ which has the expression:

$$\frac{1}{(2\pi\tau_i)^{n_i/2}} e^{-\frac{1}{2\tau_i}(\underline{y}_{T_{i-1}}^{T_i} - X_i\underline{\beta}_i)'(\underline{y}_{T_{i-1}}^{T_i} - X_i\underline{\beta}_i)}$$

Since the segments are statistical independent, the pdf for the set of observations $\underline{y}_0^n$ is :

$$f_{r,T,K}(\underline{y}_0^n | \underline{\beta}_1, \ldots, \underline{\beta}_r, \tau_1, \ldots, \tau_r) = \prod_{i=1}^{r} f_{T_{i-1},T_i,k_i}(\underline{y}_{T_{i-1}}^{T_i} | \underline{\beta}_i, \tau_i) \tag{2}$$

which depends on the sequence $T = T_0, T_1, \ldots, T_r$ of commutation times and on the sequence $K = k_1, k_2, \ldots, k_r$ of numbers of coefficients of polynomials. We rephrase the segmentation problem by using the MDL terminology: we denote $\xi = \{r, T, K\}$ the class (structure) of the model and for each ξ define the set of model parameters $\theta^{(\xi)} = \left\{\underline{\beta}_1^{(\xi)}, \underline{\beta}_2^{(\xi)}, \ldots, \underline{\beta}_r^{(\xi)}, \tau_1^{(\xi)}, \tau_2^{(\xi)}, \ldots, \tau_r^{(\xi)}\right\}$. The class of the model ξ defines a partition of the original data $\underline{y}_0^n$ in non-overlapping equivalence classes [3] (one equivalence class is associated with each segment) and due to this property the likelihood function (2) factors into a product of individual likelihood functions and the stochastic complexity [1][5][6] is obtained by summation of the stochastic complexity for each segment. We also need to consider the stochastic complexity $\mathrm{L}(\xi)$ associated to ξ.

For a given ξ a linear regression problem has to be solved for each segment i, the stochastic complexity being given by $-\log \hat{f}_i(\underline{y}_{T_{i-1}}^{T_i} | k_i)$ where $\hat{f}_i(\underline{y}_{T_{i-1}}^{T_i} | k_i)$ denotes the Normalized Maximum Likelihood (NML) density function [1][6]. In [7] two exact expressions were derived for the NML density function in the particular case of linear regression problem: the first one depends on two (hyper)parameters and the second one is parameter free. We will use in the sequel the parameter free expression given by equation (19) from [7].

Summarizing, solving the segmentation problem reduces

to the minimization of the stochastic complexity criterion:

$$\hat{\xi} = \text{argmin}_{\xi} \left\{ \text{L}(\xi) - \sum_{i=1}^{r} \log \hat{f}_i(\underline{y}_{T_{i-1}}^{T_i} | k_i) \right\} \quad (3)$$

Since the code length for the number of coefficients of polynomials $k_1, k_2, \ldots, k_r$ does not have an important contribution to the term $\text{L}(\xi)$ we will neglect its contribution. It remains that $\text{L}(\xi)$ is given only by the code length for encoding the number of segments, r, and the commutation times $T_0, T_1, \ldots, T_r$. The next section introduces several alternative expressions for $\text{L}(\xi)$.

2 THE COMPLEXITY OF THE MODEL CLASS

Before discussing the complexity of the model class we need to introduce the model of the commutation times. Let δ_t be the binary random variable which takes value 1 when t is a commutation time ($t \in \{T_0, T_1, \ldots, T_r\}$) and is 0 otherwise. We can associate to a given segmentation the binary string δ_0^n obtained by concatenating the values of δ_t at the time moments $t = 0, 1, \ldots, n-1$. We assume that the components of the string are i.i.d. and distributed according to Bernoulli law with unknown parameter q ($\delta_t = 1$ with probability q).

A natural choice in obtaining an expression for $\text{L}(\xi)$ is to first encode the number of segments and then the actual segmentation conditioned to r. If we consider that all possible values for the number of segments are equally likely, and once the number of segments is known any segmentation scenario is equally likely, it follows from the combinatorial complexity formula [3][5] that:

$$\text{L}(\xi) = \log n + \log \left(\begin{array}{c} n-1 \\ r-1 \end{array} \right) \quad (4)$$

Another approach is to code directly the binary string δ_0^n. If $\pi(q)$ is the prior for the parameter q, the stochastic complexity of the string is given by [5]:

$$\text{L}(\xi) = -\log \int_0^1 q^r (1-q)^{n-r} \pi(q) dq \quad (5)$$

When $\pi(q)$ is the uniform prior (situation which corresponds to the Laplace estimator) the expression (5) can be exactly evaluated and

$$\text{L}_L(\xi) = \log(n+1) + \log \left(\begin{array}{c} n \\ r \end{array} \right) \quad (6)$$

The difference between (4) and (6) results from considering in the Laplace estimation of the all zero string as a legal string δ_0^n.

The Krichevski-Trofimov estimator is obtained when $\pi(q)$ is the Dirichlet distribution with parameters $\left(\frac{1}{2}, \frac{1}{2}\right)$ [9]:

$$\text{L}_{KT}(\xi) = -\log \int_0^1 \frac{q^r (1-q)^{n-r}}{\pi \sqrt{(1-q)q}} dq \quad (7)$$

A relevant property of Krichevski-Trofimov estimator was established in [9]: the difference between $\text{L}_{NML}(\xi)$ and the empirical entropy is uniformly upper bounded by $\log(2\sqrt{n})$.

By using the canonical prior, we can derive the exact expression for the stochastic complexity given by NML estimator [3] $\text{L}_{NML}(\xi) = H(n,r) + \log(C_n)$ where

$$C_n = 2 + \sum_{k=1}^{n-1} \left(\begin{array}{c} n \\ k \end{array} \right) \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \quad (8)$$

and $H(n,r)$ denotes the empirical entropy $-\log \left[\left(\frac{r}{n}\right)^r \left(\frac{n-r}{n}\right)^{n-r} \right]$. We prove in Appendix that the difference between $\text{L}_{NML}(\xi)$ and the empirical entropy is uniformly upper bounded by $\log(e\sqrt{n})$ (e is base of the natural logarithm). We note that a similar bound exists for Krichevski-Trofimov estimator, but does not exist for the Laplace estimator. These bounds can be also related to the redundancy of the code. The optimality of NML code was already proved in [8].

The last possibility we consider for coding δ_0^n is to apply the two-part code [5][6]: first the estimated parameter $\hat{q}$ is coded and then the entries of the string are coded by using the already known value of $\hat{q}$. In this case the stochastic complexity is given by [3]:

$$\text{L}_{TP}(\xi) = -r \log r - (n-r) \log(n-r) + n \log n + \log(n+1) \quad (9)$$

To have an intuitive idea about how the different estimators apply to the segmentation problem, let consider the values of $\text{L}(\xi)$ when $n = 100$ and the number of segments (the number of 1's in the string δ_0^n) varies from 1 to $n-1$. We observe in Figure 1 that the shape of stochastic complexity (NML estimator) corresponds to the shape of the entropy function. In this case the stochastic complexity penalizes partitionings with greater number of segments up to $n/2$; beyond this limit the complexity decreases with the number of segments. In Figure 2 we consider the differences between the stochastic complexity for the estimators presented in this section and the empirical entropy. The code length obtained with the NML estimator is the closest to the empirical entropy.

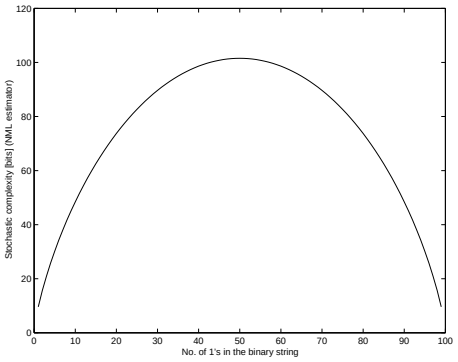

Figure 1: Dependence on the number of segments of the stochastic complexity given by the NML estimator

3 Approximating a signal with piecewise linear segments

In order to exploit the local linearity of certain types of signals we need sometimes to approximate the original signal with a "continuous broken line". Finding the best segmentation of a signal with straight lines is usually done by minimizing a criterion of the approximation errors for a given number of segments; the number of segments may also be optimized, e.g. such that errors are bellow an upper bound. It is important to note that in some applications even in the case when an approximation by straight lines achieves a low

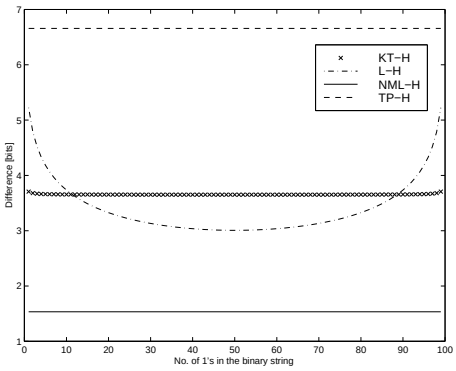

Figure 2: The differences between the stochastic complexity for the analyzed estimators and the empirical entropy

cumulative approximation, it may be objectionable because some essential features of the original signal are lost.

In the case of ECG signals, the continuous broken line approximation is used as a preliminary step before automatic signal interpretation, and also as a lossy compression method. In both cases it is crucial that the approximation does not affect the important features, i.e. location of QRS complex.

The approximation of the signal with a continuous broken line is a particular case (fixed $k_i = 2$) of the problem described in the previous section: the number of segments r and their boundaries $T_1, \ldots, T_r$ have to be estimated and for every segment i we have to estimate the gaussian noise variance τ_i and two polynomial coefficients. The use of minimum stochastic complexity as criterion is appealing since it is parameter free.

4 Experimental results

Two experiments were performed in order to evaluate the performance of MDL criterion in approximating a signal with piecewise linear segments.

In the first experiment we generate a data sequence (broken line) with length $n = 100$ where the first 50 samples are extended by periodicity, and then white gaussian noise is added. The first 50 samples are generated with the polynomials with the following coefficient vectors: $\underline{\beta}_1 = [1000, 1010]'$, $\underline{\beta}_2 = [5000, -100]'$, $\underline{\beta}_3 = [-1000, -100]'$, $\underline{\beta}_4 = [-6000, 500]'$, $\underline{\beta}_5 = [3000, 40]'$. Hence the "true segmentation" consists of $r = 10$ segments and the commutation times are: $T_0 = 0$, $T_1 = 10$, $T_2 = 18$, $T_3 = 33$, $T_4 = 42$, $T_5 = 50$, $T_6 = 61$, $T_7 = 69$, $T_8 = 84$, $T_9 = 93$, $T_{10} = 100$. The additive gaussian noise has zero mean and the variance is the same for all segments $\tau_1 = \tau_2 = \cdots = \tau_{10} = \sigma^2$. During the experiments 3 different values for noise standard deviation were considered: $\sigma = 0.01, 0.1, 1.0$. The minimization of the segmentation criteria was done for 100 different realizations for each value of noise variance. We compare the results when using the MDL, MAP and AIC criteria (the setting for the last two methods is the one used in [2]). We selected as MDL rule the expression (3) where the second term is given by equation (19) from [7] and the first term is the NML estimator as it was described in Section 2. Since both terms of expression are obtained by applying the NML technique we use the name NML (instead of MDL) in Table 1 to avoid confusion with older forms of MDL criterion.

The computation of MAP and AIC rules was done in conformity with the equations (9), respectively (16) from [2]. Since no restriction on the number of possible segments is imposed, the computational complexity of brute force evaluation of the criteria for all possible sequences $T_0, \ldots, T_{100}$ is extremely high. We use a dynamic programming scheme for the efficient evaluation of all criteria.

Table 1 shows the estimated number of segments ($\hat{r}$) for three different values of noise standard deviation (σ). The best results (perfect segmentations) are obtained with NML(MDL) rule, while the MAP and AIC grossly overestimate the number of segments when the noise variance becomes large. The result of the experiment strongly favor the use of NML algorithm for the segmentation of noisy signals.

The second experiment was performed for the piecewise linear approximation of ECG signals by using NML rule (with possible applications to signal analysis or signal compression before storing). The original ECG signal sampled at 100 Hz was split in non-overlapping frames, each frame containing 100 samples (slightly more than one period, the heart rate being 69 beats/minute). The original ECG signal contains 10000 samples, the total number of frames is 100. The dynamic programming scheme for determining the best criteria was applied independently for each frame. The linear regression term in the stochastic complexity criterion is always the same, namely NML criterion, but for the term accounting for the complexity of the model structure we alternatively tested all the estimators described in Section 2.

The result of the experiment was that for all the estimators described in Section 2 the same segmentation of the ECG signal was obtained (same number of segments and same changing points inside each frame).

This leads to the same value of the percent root-mean-square difference (PRD) measure. The PRD is a commonly used indicator for ECG segmentation and is given by $\frac{||\hat{\underline{y}}_0^n - \underline{y}_0^n||}{||\underline{y}_0^n||} \times 100$ where $||\cdot||$ denotes the Euclidean norm and $\hat{\underline{y}}_0^n$ is the broken line which approximates the observations $\underline{y}_0^n$. The PRD takes the same value for all the estimators, but we remark that coding the changing times with NML technique ensures the smallest stochastic complexity. Regarding the whole scheme as a lossy coder it results that the NML estimator guarantees the best tradeoff between code length and the PRD. Figure 3 shows the estimated number of segments and the PRD for the considered 100 frames. A very important property of the obtained segmentations: all algorithms conserve the location of QRS complex which is essential for accepting the method as a valid ECG lossy coding method. Figure 4 presents an example of how the piecewise linear approximation preserves the features of the signal.

5 Appendix

We have to prove $C_n < e\sqrt{n}$. After some computations and by using the general forms of the Abel identities [4] we obtain:

$$C_n = \frac{n!}{n^n} \sum_{k=0}^{n} \frac{n^k}{k!} < \frac{n!}{n^n} e^n \qquad (10)$$

We consider the sequence $(a_n)_{n \geq 1}$, $a_n = \frac{n! e^n}{n^n \sqrt{n}}$ and will prove that $(a_n)_{n \geq 1}$ is a decreasing sequence. Then the inequality for C_n will result from (10) and the observation $a_1 = e$.

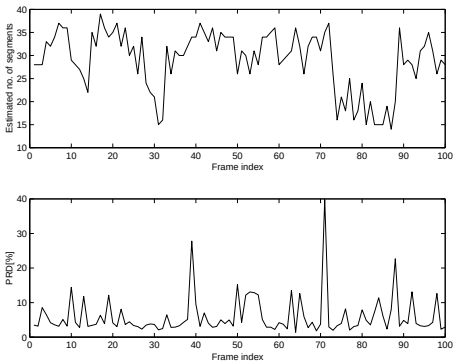

Figure 3: The estimated number of segments and the percent root-mean-square difference (PRD) for 100 frames of ECG signal

Criterion	σ	$\hat{r}$				
		10	11	[12,15]	[16,19]	≥ 20
NML	0.01	100	0	0	0	0
	0.1	100	0	0	0	0
	1.0	100	0	0	0	0
MAP	0.01	100	0	0	0	0
	0.1	37	11	43	9	0
	1.0	0	0	0	36	64
AIC	0.01	0	0	0	0	100
	0.1	0	0	0	0	100
	1.0	0	0	0	0	100

Table 1: Values of $\hat{r}$ (estimated number of segments) obtained segmenting 100 noisy realizations of a piecewise straight line by the methods using the following criteria: MDL(NML), MAP and AIC. The true r was in all cases $r = 10$.

Since $\frac{a_{n+1}}{a_n} = e\left(\frac{n}{n+1}\right)^{n+1/2}$, the monotonicity of the sequence $(a_n)_{n\geq 1}$ results from the inequality

$$\left(\frac{n+1}{n}\right)^{n+1/2} > e \tag{11}$$

Let's define $f : \Re^+ \to \Re, f(x) = \left(\frac{x+1}{x}\right)^{x+1/2}$. The derivative of $\ln[f(x)]$ is given by

$$g(x) = \ln\left(\frac{x+1}{x}\right) - \frac{2x+1}{2x(x+1)} \tag{12}$$

By applying the substitution $\alpha = \frac{1/2}{x+1/2}$, $g(x)$ becomes

$$\begin{aligned} g(\alpha) &= \ln\left(\frac{1+\alpha}{1-\alpha}\right) - \frac{2\alpha}{1-\alpha^2} \\ &= 2\left(\alpha + \frac{\alpha^3}{3} + \frac{\alpha^5}{5} + \ldots\right) - \frac{2\alpha}{1-\alpha^2} \\ &\leq 2\left(\alpha + \alpha^3 + \alpha^5 + \ldots\right) - \frac{2\alpha}{1-\alpha^2} \\ &= \frac{2\alpha}{1-\alpha^2} - \frac{2\alpha}{1-\alpha^2} = 0 \end{aligned}$$

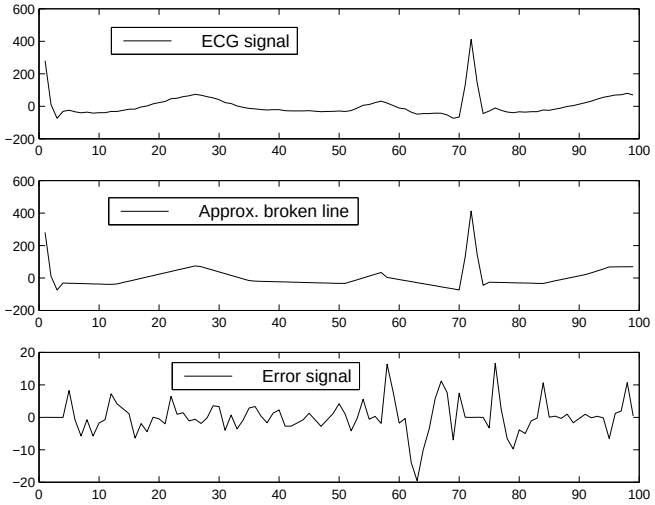

Figure 4: (top) Original ECG signal; (middle) piecewise linear segmentation; (bottom) approximation errors.

Hence $f(x)$ is a decreasing function, which implies that the sequence $(b_n)_{n\geq 1}, b_n = \left(\frac{n+1}{n}\right)^{n+1/2}$ is decreasing too. Since $\lim_{n\to\infty} b_n = e$ the inequality (11) results directly.

References

[1] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44:2743–2760, Oct. 1998.

[2] P.M. Djuric. A MAP solution to off-line segmentation of signals. In *Proc. ICASSP-94*, volume 4, pages 505–508, Adelaide, South Australia, April 1994.

[3] B.E. Dom. MDL estimation with small sample sizes including an application to the problem of segmenting binary strings using Bernoulli models. Technical Report RJ 9997, IBM, Almaden Research Center, Dec. 1995.

[4] J. Riordan. *Combinatorial identities*. John Wiley & Sons, Inc., 1968.

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

[6] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47, Jan. 1996.

[7] J. Rissanen. MDL denoising. Technical report, http://www.cs.tut.fi/∼rissanen, Dec. 1999.

[8] J. Rissanen. Strong optimality of the Normalized ML models as universal codes. Technical report, http://www.cs.tut.fi/∼rissanen, Mar. 2000.

[9] M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The Context-Tree weighting method: basic properties. *IEEE Transactions on Information Theory*, 41:653–664, May 1995.