

Piecewise Continuous Segmentation of Multidimensional Experimental Signals

Alexander G. Dmitriev
Cherepovets Higher Military
Engineering School of Radio
Electronics
Cherepovets, Russia
dag334a@fxmail.ru

Abstract

An algorithm for piecewise continuous approximation of structural experimental multidimensional signals with a previously unknown number of intervals for splitting signals into "similar" fragments is proposed. The construction of a multidimensional piecewise continuous approximating function is performed "left – to – right", which allows to use the dynamic programming method to determine the boundaries of the partition intervals. The approximation quality criterion is used, taking into account the number of data on the partition intervals and the "complexity" of the local signal models used.

Introduction

In various applications, the problem arises of analyzing the so-called structural experimental signals, considered as a time-sequential combination of simpler signals (functions), that have constant properties at the corresponding time intervals [Mot79, Kos04]. Processing of such signals in most cases is reduced to a two-stage procedure: the allocation of "similar" fragments (segmentation stage) and the subsequent construction of the description of the presented signals as a whole. The use of existing methods of signal segmentation proves to be insufficiently effective in studies under conditions of high dimensionality, limited experimental observations and an unknown

number of "similar" sites (partition intervals). In addition, the approximating function usually suffers a gap at the boundaries of "similar" sites [Dmi10, Dor84].

The aim of the work is to develop an algorithm for piecewise continuous approximation of multidimensional signals with a previously unknown number of intervals of splitting signals into "similar" fragments, which under certain conditions delivers the optimal value to the selected criterion of approximation quality.

1 Problem Statement

Let a set of signals (a multidimensional signal) $y(t) = (y^{(1)}(t), \dots, y^{(s)}(t))$ collectively characterizing the object under study, be presented for analysis. The values $y^{(i)}(t)$, $i = 1, \dots, s$ are specified at discrete points in time $t = t_1, \dots, t_N$. The criterion of approximation quality J on the sample of experimental values is chosen as:

$$J = \sum_{i=1}^s \mu_i \sum_{j=1}^r \frac{n_j}{n_j - m} \sum_{t_i \in (T_{j-1}, T_j]} [y^{(i)}(t_i) - F_j^{(i)}(t_i, \alpha_j^{(i)})]^2, \quad (1)$$

where μ_i – a priori "weight" of the signal $y^{(i)}$, n_j – the number of discrete samples on the interval $(T_{j-1}, T_j]$, $F_j^{(i)}(t_i, \alpha_j^{(i)}) = \sum_{k=1}^m \alpha_{jk}^{(i)} \varphi_k(t)$ – polynomial over a given set of basis functions $\{\varphi_k(t), k = 1, \dots, m\}$, $\alpha_j^{(i)} = (\alpha_{j1}^{(i)}, \dots, \alpha_{jm}^{(i)})$ – the vector of estimated parameters in j -th interval.

Criterion (1) uses weights μ_i ($i = 1, \dots, s$) and $n_j / (n_j - m)$ ($j = 1, \dots, r$). The introduction of weights μ_i is due to the fact that in practice the parameters often have different practical significance. Specific values μ_i are chosen for meaningful reasons, usually

Copyright c by the paper's authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). In: A. Khomonenko, B. Sokolov, K. Ivanova (eds.): Selected Papers of the Models and Methods of Information Systems Research Workshop, St. Petersburg, Russia, 4-5 Dec. 2019, published at <http://ceur-ws.org>

they are normalized so that $\sum_{i=1} \mu_i = 1$. Weights

$n_j / (n_j - m)$ are the usual normalizing coefficients that take into account the dimension of the model.

It is required to find such a partition $T = (T_0, T_1, \dots, T_r)$, $T_0 < T_1 < \dots < T_r$ (synchronous change of signals $y^{(i)}$ is assumed) of a given interval $[t_1, t_N]$, $T_0 = t_1$, $T_r = t_N$ into r intervals $(T_{j-1}, T_j]$, $j = 1, \dots, r$ (r - in general, unknown) and to determine on each of these intervals such values of the parameter vectors $\alpha_j^{(i)}$, $i = 1, \dots, s$, that the functional (1) takes a minimum value under the condition of restrictions on the continuity of approximating functions:

$$\begin{aligned} F_j^{(i)}(T_j, \alpha_j^{(i)}) &= F_{j+1}^{(i)}(T_j, \alpha_{j+1}^{(i)}), \\ (j &= 1, \dots, r); \quad (i = 1, \dots, s). \end{aligned} \quad (2)$$

2 Algorithm

Denote by the $\varepsilon^{(i)}(T_{j-1}, T_j)$ error of approximation of the i -th signal on the interval $(T_{j-1}, T_j]$, calculated by the method of least squares. Then criterion (1) will take the form:

$$J = \sum_{i=1}^s \mu_i \sum_{j=1}^r \frac{n_j}{n_j - m} \varepsilon^{(i)}(T_{j-1}, T_j) \quad (3)$$

Let's make the following assumption. Borders T_j ($j = 1, \dots, r-1$), partitions $T = (T_0, T_1, \dots, T_r)$ and corresponding local approximating functions on each interval of partition will be found from "left to right". Let the position of the boundary T_1 be determined, find the local approximating functions $F_1^{(i)}(t, \alpha_1^{(i)})$, delivering the minimum $\varepsilon^{(i)}(T_1, T_2)$, $i = 1, \dots, s$. Next, we fix the position of the boundary T_1 , find the boundary T_2 and the corresponding local approximating functions $F_2^{(i)}(t, \alpha_2^{(i)})$, delivering a minimum of $\varepsilon^{(i)}(T_1, T_2)$ ($i = 1, \dots, s$) with a restriction on continuity at the boundary T_1 :

$$F_1^{(i)}(T_1, \alpha_1^{(i)}) = F_2^{(i)}(T_1, \alpha_2^{(i)}), \quad i = 1, \dots, s.$$

Fix T_2 , etc.

Under this assumption criterion (3) with constraints (2) has the following property. Let be T_j^* some fixed position of the right boundary of the j -th interval. Then the bounds T_1^*, \dots, T_{j-1}^* , obtained by minimizing (3) only over the bounds T_1, \dots, T_{j-1} are independent of the boundary values T_{j+1}, \dots, T_{r-1} . Indeed, the functional (3) can be represented as the sum of two nonnegative quantities: $J = J_a + J_b$, where

$$J_a = J_a(T_1, \dots, T_{j-1} / T_j^*) = \left\{ \sum_{k=1}^j \frac{n_k}{n_k - m} \sum_{i=1}^s \mu_i \varepsilon^{(i)}(T_{k-1}, T_k) \right\},$$

$$T_j = T_j^*,$$

$$J_b = J_b(T_{j+1}, \dots, T_{r-1} / T_j^*) = \left\{ \sum_{k=j+1}^r \frac{n_k}{n_k - m} \sum_{i=1}^s \mu_i \varepsilon^{(i)}(T_{k-1}, T_k) \right\},$$

$$T_j = T_j^*.$$

But then, obviously, $\arg \min_{T_1, \dots, T_{j-1}} J = \arg \min_{T_1, \dots, T_{j-1}} J_a$.

It follows from this property that if T_j^* is the optimal position of the j -th boundary, then the boundaries T_1^*, \dots, T_{j-1}^* , obtained by minimizing J by T_1, \dots, T_{j-1} , are also optimal. The considered property of criterion (3) makes it possible to use the dynamic programming procedure [Bel69] to determine the optimal boundaries of intervals.

Let the number of intervals be equal to r_0 .

The following recursive algorithm finds the partition and local approximating functions, delivering the optimal value to the functional (1) (under the above assumption).

First the functions $J_j(T_j)$ ($j = 1, 2, \dots, r_0 - 1$) are tabulated sequentially, where

$$J_1(T_1) = \frac{n_1}{n_1 - m} \sum_{i=1}^s \mu_i \varepsilon^{(i)}(T_0, T_1), \quad T_1 = t_p, \dots, t_{N-(r_0-1)p};$$

$$J_j(T_j) = \min_{\substack{T_{j-1} = t_{(j-1)p}, \\ \dots, t_{T_j-p}}} \left\{ J_{j-1}(T_{j-1}) + \frac{n_j}{n_j - m} \sum_{i=1}^s \mu_i \varepsilon^{(i)}(T_{j-1}, T_j) \right\},$$

$$T_j = t_{j \cdot p}, \dots, t_{N - (r_0 - j)p}, \quad j = 2, \dots, r_0 - 1;$$

T'_j – the reference number corresponding to the boundary T_j , p is the specified minimum allowed number of samples on the partition interval. Approximating functions on adjacent intervals are constructed taking into account the conditions for continuity (2).

At the same time, the values $M_{j-1}(T_j)$, $T_j = t_{j \cdot p}, \dots, t_{N - (r_0 - j)p}$, $j = 2, \dots, r_0 - 1$, – the values of the optimal positions of the boundaries T_{j-1} for each T_j , – are stored. Next, the optimal boundaries of the intervals are determined:

$$T_{r_0-1}^* = \arg \min_{T_{r_0-1} = t_{(r_0-2)p}, \dots, t_{N-p}} \left\{ \begin{array}{l} J_{r_0-1}(T_{r_0-1}) + \\ + \frac{n_{r_0}}{n_{r_0} - m} \sum_{i=1}^s \mu_i \varepsilon^{(i)}(T_{r_0-1}, t_N) \end{array} \right\}, \quad (4)$$

$$T_{r_0-2}^* = M_{r_0-2}(T_{r_0-1}^*), \dots, T_1^* = M_1(T_2^*).$$

The extreme nature of the dependence J on r is used to find the partition $T^H = (T_0, T_1^H, \dots, T_{r_H-1}^H, T_r)$ at an unknown number of intervals. Indeed, with an increase r , on the one hand, there is an increase in weights $n_j / (n_j - m + 1)$ due to a decrease in the average n_j , which, other things being equal, leads to an increase in the criterion (1). On the other hand, as the number of intervals increases, the usual quadratic residual decreases, resulting in a decrease in (1). The simultaneous action of these factors leads to the fact that the functional (1) reaches its minimum value at some intermediate (not boundary) value r_H .

For determination r_H it is possible to use the approach offered in work [Dmi10]. First, the minimum values of the functional (1) $J_j(t_N)$, $j = 2, \dots, r_{\max}$ are calculated. At the same time, the values $M_{j-1}(T_j)$ of the optimal positions of the boundaries T_{j-1} for each T_j are stored. Next, the number of intervals is selected and the optimal boundaries are determined. As r_H the smallest number of intervals r is chosen, at which $J_r(t_N)$ takes the minimum value. Optimal

bounds $T_1^H, \dots, T_{r_H-1}^H$ are determined from expressions analogous to (4):

$$T_{r_H-1}^H = M_{r_H-1}(t_N), \quad T_{r_H-2}^H = M_{r_H-2}(T_{r_H-1}^H), \\ \dots, \quad T_1^H = M_1(T_2^H).$$

The value r_{\max} is chosen for substantive or Statistical reasons. In particular, as r_{\max} can be used the value $[N/p]$, where $[x]$ is the integer part x .

Modeling

To check the efficiency of the algorithm, its modeling was carried out on special multidimensional signals. The developed program made it possible to obtain a multidimensional signal with specified properties. The relevant procedure is as follows.

The vector function is considered on the interval: $[1, N]$:

$$y(t) = (y^{(1)}(t), \dots, y^{(s)}(t)),$$

where

$$y^{(i)}(t) = \sum_{j=1}^{r^*} \varepsilon_j \left\{ \frac{y_j^{(i)} \cdot T_{j-1} - y_{j-1}^{(i)} \cdot T_j + (y_{j-1}^{(i)} - y_j^{(i)})t}{T_{j-1} - T_j} \right\} + b \cdot \xi_t \quad (5)$$

$$(i = 1, \dots, s)$$

is a superposition of piecewise linear continuous function and independent Gaussian interference with zero mean and dispersion b^2 ; $T_0 = 1$, $T_{r^*} = N$, T_j , $j = 1, \dots, r^* - 1$ – nodal points of piecewise linear vector function; $y_j^{(i)}$, $j = 0, \dots, r^*$ – values of this function in points T_j ; ε_j – characteristic function equal to one, if $t \in (T_{j-1}, T_j]$, and equal to zero, if $t \notin (T_{j-1}, T_j]$.

The nodal points T_j , the values of piecewise linear functions $y_j^{(i)}$ at these points, the number of intervals r^* are determined randomly using a random number generator by the following procedure:

1. $j = 0$; $T_0 = 1$; $y_0^{(i)} = \zeta_{0i}$, $i = 1, \dots, S$.

2. If $N - T_j < p$, then move to 4. Otherwise - we pass to 3.

3. $j = j + 1$; $T_j = T_{j-1} + [\beta_j c + p]$, ($[x]$ is the integer part x), $y_j^{(i)} = \zeta_{ji} \cdot d$, $i = 1, \dots, s$, then move to 2.

4. $T_j = N$; $r^* = j$.

ζ_{ji}, β_j - random, independent, uniformly distributed, respectively, at intervals $[-1; 1]$ and $[0; 1]$ values; p, d, c, N - pre-selected parameters of the algorithm.

Next, the values $y^{(i)}(t)$, ($t = 1, 2, \dots, N$), ($i = 1, \dots, s$) of the vector function (5) are calculated.

Linear functions were used as local approximating functions:

$$F_j^{(i)}(t, \alpha_j^{(i)}) = \alpha_{j1}^{(i)} + \alpha_{j2}^{(i)} \cdot t, \quad i = 1, \dots, s, \quad j = 1, \dots, r^*.$$

In the studies, the experimental material contained three groups of three-component multidimensional signals, obtained using the procedure described above. The first group consisted of multidimensional signals without noise, the second and third, respectively, with average (or low) (dispersion $b^2 \in [0, 01 - 0, 1]$) and increased noise level ($b^2 \in [0, 1 - 0, 3]$). As expected, for no-noise the algorithm found the desired number of intervals, partitioning, and approximating functions without error.

Typical dependences of the criterion J on r for the second signal group are shown in figures 1

(the \uparrow sign in the figures indicates the actual number of intervals).

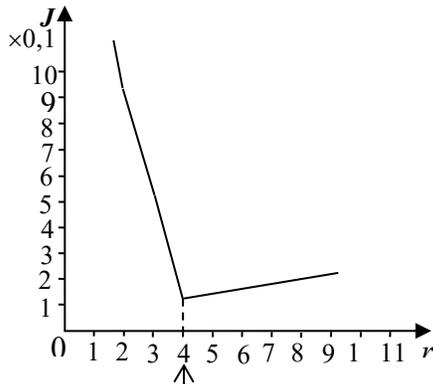


Figure 1: Average noise level

For signals with an average noise level, the minimum functional, as a rule, falls on the desired number of intervals.

Typical dependences of the criterion J on r for the third signal group are shown in figures 2.

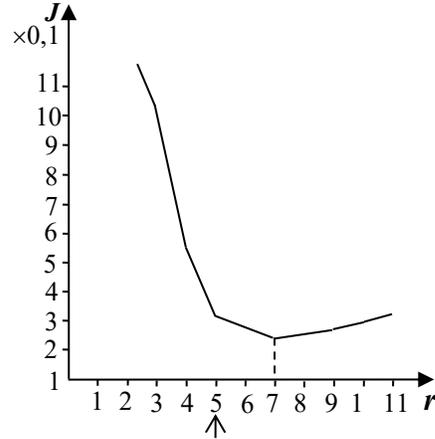


Figure 2: Increased noise level

For signals with an increased noise level, the discrepancy between the optimal and the desired number of intervals was more often manifested. For signals with high noise levels were more likely to be a mismatch is found and the specified numbers of intervals.

For both groups of signals, this discrepancy was observed when there were adjacent intervals in the multidimensional signal, at which the difference in the "behavior" of the signal was insignificant, and, as a rule, such intervals contained a small number of samples.

It can also be noted, that for the signals of the second group, the reduction section of the graph of the dependence of J on r jumps into the increase section, and for signals with an increased noise level, such a transition occurs smoothly. The latter can be used to construct optimization procedures that perform a limited search over r .

Conclusion

Thus, the proposed approach of constructing a piecewise continuous approximating function "left - to - right" allows to apply the dynamic programming

method to determine the boundaries of the partition intervals. Using the extreme behavior of the approximation quality criterion, the number of partition intervals is determined.

References

- [Mot79] V. V. Mottl, I. B. Muchnik. *Hidden Markov models in structural analysis of signals*. M.: Fizmatlit, 1999.
- [Dmi10] A. G. Dmitriev. The Algorithm of optimal structural approximation of experimental multidimensional signals. *Naukoemkie tehnologii*, No 9: 31-35, 2010.
- [Kos04] A. A. Kostin, O. V. Krasotkina, M. V. Markov, V. V. Mottl, I. B. Muchnik. Dynamic programming algorithms for analysis of non-stationary signals. *Computational Mathematics and Mathematical Physics*, V 44, No 1: 62–77, 2004.
- [Dor84] A. A. Dorofeyuk, A. G. Dmitriev. Methods for piecewise approximation of multidimensional curves. *Automatics and telemekhanics*, No 12: 101-108, 1984.
- [Bel69] R. Bellman, S. Dreyfus. *Applied problems of dynamic programming*. M: Nauka, 1969.