Preprints of Control Conference Africa 2024
Balaclava, Mauritius, 16-17 September 2024

Fast and Tunable Segmentation and Trend Extraction of Process Data with Constrained Polynomial Fitting ⋆

Ju Liu [*,**] **Dexian Huang** [*] **Chao Shang** [*]

[*] *Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: liu-j21@mails.tsinghua.edu.cn; huangdx@tsinghua.edu.cn; c-shang@tsinghua.edu.cn;).*
[**] *Luoyang Electronic Equipment Test Center, Henan 471000, China*

Abstract: Generic segmentation algorithms are based upon piesewise polynomial fitting, where the number of data segments is difficult to select and expensive computations are entailed in algorithms based on dynamic programming (DP). To address these issues, we propose a novel data segmentation and trend extraction method. Firstly, we set constraints on the error of each segment. Secondly, a breadth-first search approach with branch pruning (BP) strategies is established to solve for the optimal segmentation points efficiently. Comprehensive case studies show that the algorithm has good accuracy and efficiency.

Keywords: Data segmentation, trend extraction, breadth-first, polynomial fitting.

1. INTRODUCTION

The key step of segmentation and trend extraction is finding segmentation points. In existing works, the methods to find segmentation points can be divided into two cases: statistical methods and global optimization-based methods. The former method picks out the segmentation points through a statistical test . The latter method searches for the segmentation points by solving optimization problems, the objective function usually contains residuals and penalty. Both of the methods mentioned above have been successfully used in real-world processes. However, the two methods both depend on parameter settings and the method based on global optimization has large computation when dealing with long data segments.

To address these limitations, we propose a fast and tunable segmentation and trend extraction method based on constrained polynomial fitting. We use *the fewest segments* to ensure that each segment satisfies the constraints. To search for the optimal segmentation points, we proposed a tree search method with efficient branch pruning strategies. Compared with the existing algorithms, the proposed algrithm has better accuracy and efficiency.

2. PROBLEM DESCRIPTION

Consider a scalar-valued time series $\mathbf{y} = [y_{t_1}, y_{t_2}, \cdots, y_{t_n}]^\top$ of length n, where $t_1, \cdots, t_n$ are the sampling instances. To capture the trend of $\mathbf{y}$ and remove the noise, one can approximate it using a q-order polynomial:

$$\hat{y}_t = \beta_0 + \beta_1 t + \cdots + \beta_q t^q, \tag{1}$$

where $\hat{y}_t$ is the approximate value of y_t, $t = t_1, \cdots, t_n$, and $\beta_0, \beta_1, \cdots, \beta_q$ are polynomial coefficients. One can further compactly express (1) as:

$$\hat{\mathbf{y}} = [\hat{y}_{t_1}, \hat{y}_{t_2}, \cdots, \hat{y}_{t_n}]^\top = \mathbf{T}\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_q]^\top$, $\mathbf{T}$ is the time index matrix with entries $T_{i,j} = t_i^j$, $1 \le i \le n$, $1 \le j \le q$. The goal of piesewise polynomial fitting is to minimize the total cost of M segments:

$$f(M) = \min_{\substack{s_1, s_2, \cdots, s_{M-1} \\ \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_M}} \sum_{m=1}^{M} V(\hat{\mathbf{y}}_m, \mathbf{y}_m) + \lambda g(M), , \tag{3}$$

where $V(\hat{\mathbf{y}}_m, \mathbf{y}_m)$ is the cost of the segment $\mathbf{y}_m$. It can be usually selected as $\|\hat{\mathbf{y}}_m - \mathbf{y}_m\|^2$, $g(M)$ is the penalty on M and monotonically increasing, λ is the penalty coefficient. There exist criteria that determine the number of segments, such as Schwarz (Chung-Bow and Lee (1995)). For a given number M of segments, the complexity is $\mathcal{O}(Mn^2)$ (Bellman (1961)). These criteria need to consider all the possible number of segments, which will further increase the complexity. A special penalty is $g(M) = M$, and the complexity can be reduced to $\mathcal{O}(n^2)$ by DP (Jackson et al. (2005)). The disadvantage of these methods is that the penalty coefficient λ is difficult to select, and the accuracy of each segment cannot be guaranteed.

3. FAST AND TUNABLE SEGMENTATION AND TREND EXTRACTION METHOD

In this paper, we propose to use *the fewest segments* to fit the entire trajectory while making the approximation accuracy sufficiently high. That is, the penalty coefficient λ in (3) is set large enough and the cost $V(\mathbf{y}_m)$ is constrained by approximation accuracy. $V(\mathbf{y}_m)$ is defined as:

$$V(\hat{\mathbf{y}}_m, \mathbf{y}_m) = \begin{cases} \min\limits_{\boldsymbol{\beta}_m} \|\hat{\mathbf{y}}_m - \mathbf{y}_m\|^2 \\ \text{s.t. } \hat{\mathbf{y}}_m \in \mathcal{Y}_m, \end{cases} \tag{4}$$

⋆ This work was supported by National Natural Science Foundation of China under Grant 62373211.

where $\hat{\mathbf{y}}_m \in \mathcal{Y}_m$ indicates that the approximation accuracy of $\hat{\mathbf{y}}_m$ is "satisfactory" in some sense. We define $\hat{\mathbf{y}}_m \in \mathcal{Y}_m$ as:

$$\|\hat{\mathbf{y}}_m - \mathbf{y}_m\|_2 \leq \epsilon_{\mathrm{N}} \mathcal{S}(\mathbf{y}_m), \tag{5a}$$

$$\|\hat{\mathbf{y}}_m - \mathbf{y}_m\|_\infty \leq \epsilon_{\mathrm{S}} \mathcal{S}(\mathbf{y}_m),\ j = 1, \cdots, n_m \tag{5b}$$

where $\mathcal{S}(\mathbf{y})$ is the *smoothness index* of $\mathbf{y}$:

$$\mathcal{S}(\mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n-1} (y_{i+1} - y_i). \tag{6}$$

Obviously, the constraint (5a) is just on the objective function of (4). Thus, we can set only the constraint (5b) on $\mathbf{y}_m$ when solving problem in (4) and then test whether the solution meets the constraint (5a). If the problem in (4) is infeasible, then $V(\hat{\mathbf{y}}_m, \mathbf{y}_m) = +\infty$. Thus, the problem (3) can be described as follows:

$$f(M) = \begin{cases} \min\limits_{\substack{s_1, s_2, \cdots, s_{M-1} \\ \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_M}} \sum\limits_{m=1}^{M} V(\hat{\mathbf{y}}_m, \mathbf{y}_m) \\ \text{s.t.} \qquad \hat{\mathbf{y}}_m \in \mathcal{Y}_m,\ m = 1, \cdots, M. \end{cases} \tag{7}$$

Because λ is large enough, so the optimal number of segments is determined as:

$$M^* = \min \{M \,|f(M) < +\infty\}, \tag{8}$$

Next we delve into the efficient solution to (7). We design an tree method to find optimal segmentation points. Indeed, one can establish a one-to-one correspondence between an admissible set of segmentation points $\{s_0, s_1, \cdots, s_m\}$ and a set of *nodes* in a hierarchical tree, thereby defining a *path* $P_m = \{s_0, s_1, \cdots, s_m\}$ from the root node to a particular node in the mth level. The root node represents the first point of data trajectory.

We explore all candidate nodes level by level and evaluate the feasibility of related paths, by following the spirit of breadth-first search (BFS). The algorithm will terminates in the mth level of the tree if there exists a feasible path P_m with $s_m = n$.

Next we propose two branch pruning (BP) strategies to screen out suboptimal paths. to avoid unnecessary computations. For convenience, two strategies are termed as BP-I and BP-II. BP-I is essentially equal to the Bellman principle of optimality in DP: if there exists several feasible paths with the same endpoint, then we choose the one with the lowest cost $f(m)$.

In BP-II, if there exists several feasible paths with the same endpoint and they have different lengths, then it suffices to consider the shortest path and longer paths with more nodes are suboptimal. More formally, we assume that:

$$\begin{cases} P_{m_1} = \{1, s_1 \cdots, s_{m_1}\}, \\ P_{m_2} = \{1, s'_1 \cdots, s'_{m_2}\}, \\ s_{m_1} = s'_{m_2}, \\ m_1 > m_2, \end{cases} \tag{9}$$

then P_{m_1} is suboptimal. The reason is that the penalty coefficient λ in (3) is set large enough.

4. CASE STUDIES

The proposed method is tested in process industrial data and compared with global polynomial fit-based trend extraction (GPTE) algorithm. The experimental data comes from a heating furnace data, we set $\epsilon_{\mathrm{N}} = 1$ and $\epsilon_{\mathrm{S}} = 3$. In GPTE algorithm, the optimal number M^* of segments is determined by the bilateral criterion (Zhou et al. (2017)). When we set M_{max} to 10, 20 or 50, we get the same $M^* = 4$. The trend extraction results and computational time of the two algorithms are shown in Fig. 1 and Table. **??**. It can be seen that the trend extraction results of

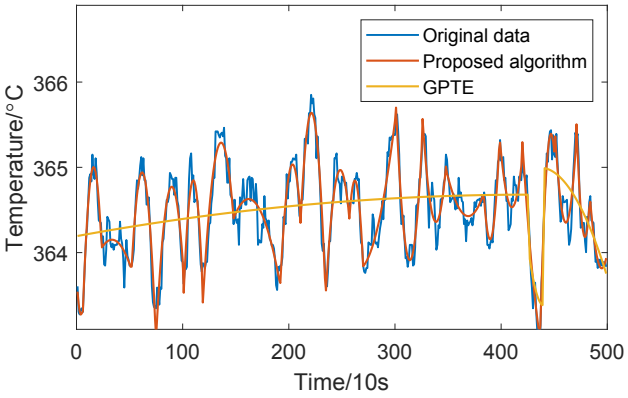

Fig. 1. Trend extraction results of the two algorithms

the proposed method are more rational, and the computational time of the proposed method is much less.

Algorithm	Parameters	Computational time
GPTE	$M_{max} = 10$	15s
	$M_{max} = 20$	32s
	$M_{max} = 50$	1min25s
Proposed algorithm	$\epsilon_{\mathrm{N}} = 1$	8s
	$\epsilon_{\mathrm{N}} = 0.8$	9s
	$\epsilon_{\mathrm{N}} = 0.6$	10s

5. CONCLUDING REMARKS

In this work, we proposed a fast and tunable tend extraction method based on polynomial fitting and tree search. We use the fewest segments to ensure that each segment satisfies the constraints on accuracy. Compared with the existing algorithm, the proposed algorithm has better accuracy and efficiency.

REFERENCES

Bellman, R.E. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*.

Chung-Bow and Lee (1995). Estimating the number of change points in a sequence of independent normal random variable. *Statistics & Probability Letters*.

Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T.T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2), 105–108.

Zhou, B., Ye, H., Zhang, H., and Li, M. (2017). A new qualitative trend analysis algorithm based on global polynomial fit. *AIChE Journal*, 63(8), 3374–3383.