Chemometrics and Intelligent Laboratory Systems 159 (2016) 138–150

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Constrained selective dynamic time warping of trajectories in three dimensional batch data

Bo Lu[a,*], Shu Xu[c], John Stuber[b], Thomas F. Edgar[c]

[a] Analytical Technology Center, The Dow Chemical Company, Freeport, TX, USA
[b] Smart Manufacturing Systems Group, VLSIP Technologies Inc., Richardson, TX, USA
[c] McKetta Department of Chemical Engineering, The University of Texas at Austin, Austin, TX, USA

ARTICLE INFO

*Keywords:*
Dynamic time warping
Trajectory alignment
Batch modeling
Three dimensional data alignment
Trajectory synchronization
Data-driven modeling

ABSTRACT

Three dimensional data structures such as batch process data or infra-red spectral measurements usually contain inconsistent trajectories of various durations and quality. In the case of batch process data, most modeling methods require the data from all batches to be of same duration. For spectral data, peaks might be shifted from one sample to another due to unaccounted sources of variation. These inconsistencies are usually resolved through trajectory alignment (or synchronization) methods. In this paper, we first review the deficiencies of existing approaches. Next, a Constrained selective Derivative Dynamic Time Warping (CsDTW) method is proposed to perform automatic alignment of trajectories. Different from conventional methods, CsDTW preserves key features that characterizes the batch and only apply warping to regions of least impact to trajectory characterization. The proposed warping technique is applied to both industrial and simulated datasets to demonstrate its effectiveness.

## 1. Introduction

Batch processes are commonly used in manufacturing of specialty chemicals, pharmaceutical, and agricultural products. However, quality measurements of the product usually come with considerable time delay. There has been a tremendous amount of effort and demonstrated success in building and deploying data-driven inferential sensors that could predict key quality measurements to improve batch process reliability and performance [1–6]. However, a challenge frequently encountered in these modeling efforts is the need for pre-processing of batch data [7]. Data collected from batch processes are inconsistent in duration and across different variables and batches. In addition, time series data in general also experiences analogous problems when comparisons are required. In these cases, it is desirable to make sure the important trajectory features in the data (rise, overshoot, drop-off, peaks, valleys) are aligned. Fig. 1 illustrates the goal of trajectory alignment and its impact on subsequent batch process modeling. In batch process modeling, it is desirable to not only align geometric trajectory features (rise, decay, peaks and valleys), but also event boundaries across different phases. Proper alignment not only improves the performance of modeling techniques, but also the interpretability of the resulting model. In addition to modeling, dynamic time warping methods have also been used in fault detection [8] and root-cause analysis of batch processes [9] (Fig. 2).

### 1.1. Overview of existing trajectory alignment techniques

There are many methods in practice that aim to synchronize trajectories of that are of different durations. These methods differ in computational complexity and also their alignment objective. Table 1 provides a brief summary of the advantages and disadvantages of each method.

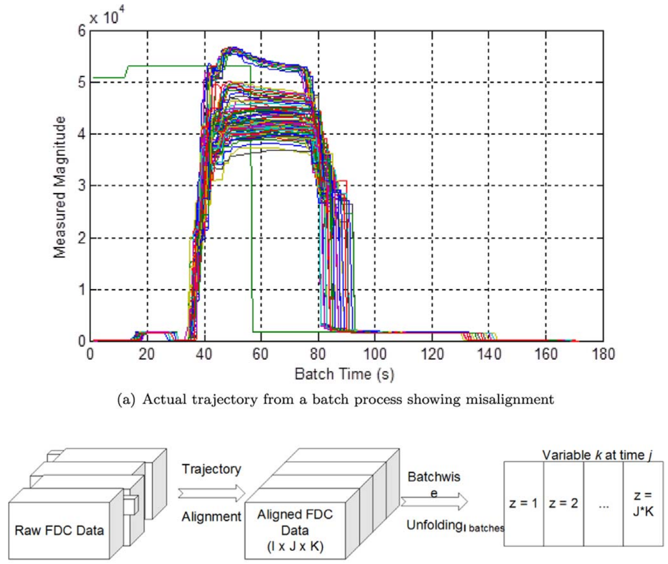
There is no alignment method that is universally superior. It is often necessary to attempt multiple alignment solutions to determine the most appropriate method for the given problem. Below, we briefly introduce the key concepts and the mechanisms behind each alignment technique.

#### 1.1.1. Truncation and padding

In truncation and padding, the alignment technique is very simple and is based mostly on intuition. The assumption made in truncation and padding is that the initial and the end of the time series data are not as important as the middle section of the data. When this assumption is valid, the longer time series trajectory can be simply shortened by removing the head or the tail of the data series. For shorter trajectories, average value from the head or tail of the data

* Corresponding author.
*E-mail address:* blu2@dow.com (B. Lu).

http://dx.doi.org/10.1016/j.chemolab.2016.10.005
Received 12 July 2016; Received in revised form 2 October 2016; Accepted 7 October 2016
Available online 19 October 2016
0169-7439/




(a) Actual trajectory from a batch process showing misalignment

(b) Schematic showing the trajectory alignment workflow

**Fig. 1.** Motivation for performing batch trajectory alignment. (a) Actual trajectory from a batch process showing misalignment (b) Schematic showing the trajectory alignment workflow.

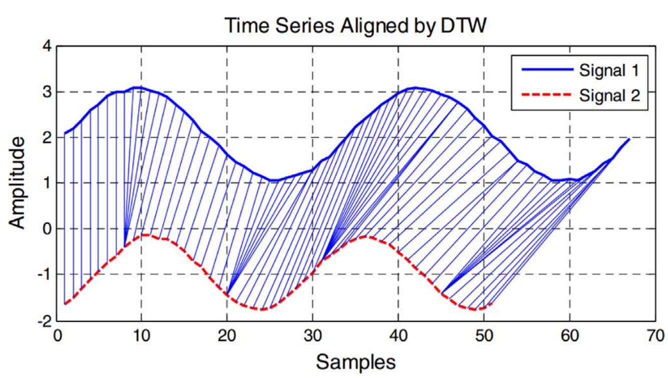



**Fig. 2.** Example solution of the Dynamic Time Warping Problem, figure taken from [15].

series is calculated and then padded on to extend the length of the trajectory being aligned.

This method guarantees that the processed trajectories are of equal length. However, this method does not ensure that the dynamic features within the time series data are properly aligned. As a result, its fidelity is often not as high as other advanced methods. However, due to its simplicity and ease of use, simple truncation or padding could serve as an effective first-pass in assessing the nature of the batch dataset being analyzed.

Lastly, truncation and padding offers the minimal amount of information loss and distortion to the given trajectory, and preserves all the dynamic features provided that the features of interest are not in the truncation region (head or tail region of the trajectory).

### *1.1.2. Linear time scaling*

Linear time scaling (LTS) is another popular alignment technique that is easy to execute and very efficient. This alignment method has been made available in many commercial data analysis software packages. In linear time scaling, a batch maturity variable is chosen from the available measurements. This variable is required to be monotonic and should ideally be indicative of the progress of a batch process. For example, if an accumulation reaction takes place inside a CSTR reactor with no outlet stream, then the level of the reactor could be an indicator of the progress of reaction. Since the level of the tank will always start empty and increases until it is full, each level reading

**Table 1**
Summary and assessment the warping methods for trajectory synchronization.

| Method | Description | Advantages | Disadvantages | Speed |
|---|---|---|---|---|
| Truncation and Padding | Delete longer time series or pad shorter time series data to the same length based on a golden reference trajectory | Easy-to-implement | Does not explicitly align features | Fast |
| Linear Time Scaling | Use a monotonic batch evolution indicator variable (percent completion, product concentration, tank level) and stretch or shrink linearly to ensure every trajectory is of the same length | Straightforward | Performance dependent on good indicator variable | Fast |
| Dynamic Time Warping | Solves dynamic programming problem to find the most similar path between two trajectories | Explicit feature alignment | Information loss, trajectory distortion, implementation challenge | Medium |
| Correlation Optimiz-ed Warping | Solves dynamic programming problem with sub-optimization routines to maximize correlation between trjaectory | Explicit feature alignment, less distortion than DTW | Information loss, complex, implementation challenge | Slow |

corresponds with an unique one-to-one mapping towards the progress of that particular batch.

To perform alignment, the indicator variable **p** is discretized from its initial value to the final value:

$$\delta p = \frac{\max(\mathbf{p}) - \min(\mathbf{p})}{K-1} p_0 = \min(\mathbf{p}) p_{k+1} = p_k + \delta p p_K = \max(\mathbf{p})$$

The number of discretization points ($K$) is a controllable parameter, but is usually set to the average number of samples in each batch observation. At each discretized value of the indicator variable $p_k$, linear interpolation is then performed for every measurement $x_j$.

$$x_{j,k} = x_j^a + (x_j^b - x_j^a) * \frac{p_k - p^a}{p^b - p^a} \tag{1}$$

where superscripts $a$ and $b$ represents the nearest boundary indices that encloses the value $p_k$. This results in $K$ number of measurement vectors $\mathbf{x}_k$ for $k \in (1, K)$. Since every batch will be interpolated to these $K$ samples after LTS, the trajectories will be aligned and ready for subsequent analysis steps.

The main advantage of LTS is that this is a very efficient technique to align multiple trajectories at the same time. Provided that an accurate linear batch progress indicator variable can be found (meaning that the value of this variable varies linearly from the start to the end). The main drawback of LTS method is the reliance of a maturity variable. Since this variable is used essentially to replace "time" as an alternative batch progress indicator, improper selection or unavailability of such a variable greatly affects the result of LTS method. However, ways to circumvent this problem using within batch PLS models have been reported [10].