Journal of Process Control 21 (2011) 1370–1377

Contents lists available at ScienceDirect

Journal of Process Control

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

Experiences in batch trajectory alignment for pharmaceutical process improvement through multivariate latent variable modelling

Salvador García-Muñoz [a,*], Mark Polizzi [a], Andrew Prpich [a], Cathal Strain [b], Adam Lalonde [b], Vilmary Negron [c]

[a] Pfizer Global Research & Development, Groton, CT 06340, USA
[b] Pfizer Global Manufacturing, Groton, CT 06340, USA
[c] Pfizer Global Manufacturing, Barceloneta, PR, United States

ARTICLE INFO

Article history:
Received 1 December 2010
Received in revised form 26 July 2011
Accepted 26 July 2011
Available online 15 September 2011

Keywords:
Batch process
Multivariate monitoring
Multi-way PCA
Multi-way PLS
Alignment

ABSTRACT

The primary objective of batch data as trajectory alignment (or synchronization) is to standardize the data sampling per batch according to the evolution of the process, and secondarily to homogenize the samples per run. The use of an indicator variable performs both objectives well. Two examples from the pharmaceutical sector are discussed to illustrate the different ways to deal with uneven samples across batches and across variables in the same batch. Since trajectory alignment requires large time investment, a simple triage approach is proposed to assess the need to analyze the dynamics of a given process and hence perform alignment. The presented examples are representative of a broad variety of batch processes that are operated by recipe in the pharmaceutical sector. In our experience, the variables associated with the automation triggers in these recipes are the best indicator variables to use for alignment. This is due to (i) the fact that the trigger variables are easy to identify from the automation of the recipe, (ii) operators are familiar with these, (iii) the target values for triggers are known a priori and hence the resulting alignment scheme can be performed in real-time for monitoring applications and (iv) it makes the monitoring scheme easy to understand and justify around the design-space since the design-space may originally be defined in terms of the trigger variables for each phase of the batch.

1. Introduction

The application of multivariate latent variable models to analyze batch processes has been widely studied and discussed in literature in applications ranging from the analysis and troubleshooting of the process using historical data from the network of sensors installed in the process; process monitoring and fault detection for continuous quality assurance; process operation design and optimization; and control. These techniques have been successfully implemented in industrial settings, with some applications available in the public literature and excellent reviews written on the topic [1,2].

A batch recipe will commonly be executed by a series of instructions that trigger the available actuators based on process targets (temperature, weights, pressured, etc.) and will often have uneven time-length. Due to this uneven time duration across batches a key exercise in performing any statistical analysis of batch data is the need to align (or synchronize) the samples taken throughout the run, for all batches available. This is necessary so that sample *i* (across batches) correspond to the same state of evolution for a given process variable (i.e., temperature from the heating phase for batch *A* should not be contrasted with the temperature during the reactive phase for batch *B*). Nomikos and MacGregor [3] identified the indicator variable approach to synchronize the data. Kassidas et al. [4] later proposed to use Dynamic Time Warping when there was no other observation of the evolution of the batch.

This work presents our experience in dealing with these situations with two examples representative of those in the pharmaceutical sector. We also comment on the expectations of a batch alignment exercise from a practical perspective and finally present a triage method to assess the potential impact of the dynamics of a process onto the final product quality; and hence determine the need to invest the necessary time and effort to align the data, or not.

2. 2D Multi-way methods and process dynamics

Data sampled from a dynamic system will contain samples of variables as they change with time. The data can then be arranged in any number of ways, depending on the model structure to be used (data is only a set of numbers with some contextual

* Corresponding author. Tel.: +1 860 715 05 78.
E-mail address: sal.garcia@pfizer.com (S. García-Muñoz).

0959-1524/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2011.07.013

relationship; any structural arrangement is artificially imposed). For example, in the parameter estimation of an autoregressive with exogenous input (ARX) model, the time collected samples will be lagged depending on the order of the model. An incorrect ARX model can be built by simply assuming the incorrect order. In such case the inaccuracy of the model is not due to the ARX general structure: the problem is the incorrect order of the model! The same analogy can be applied to the application of Principal Component Analysis (PCA) on dynamic data. Numerous authors criticize PCA as unable to capture the dynamics of a process and some authors propose the inclusion of lags as an "improvement to PCA" to capture dynamics. While the approach is certainly valid -including all possible lags of batch data was the original proposal by MacGregor and Nomikos [5] – it is incorrect to state that the problem is the PCA method. The real problem is the arrangement of the data.

In a PCA model of the batch data, rearranged in a 2D matrix of $I \times (J \times K)$ dimensions (where I is the number of batches, J is the number of variables sampled during the batch and K the number of samples taken during the batch) there is one strong assumption, and that is that all the elements of a column in this matrix corresponds to a variable sampled at the same state of evolution of the batch for all batches in the data set. The same type of assumption is done if the data is first unfolded into an $(I \times K) \times J$ matrix to perform the dimension reduction to later refold the scores into an $I \times K$ matrix for each principal component in order to do establish confidence intervals and perform monitoring [6]; the assumption is that all the elements in each column of the $I \times K$ matrix corresponds to the same state of evolution for each batch.

This correspondence was discussed in early papers by Nomikos et al., who proposed a simple procedure to ensure all variables were sampled at the same state of evolution of the process: the use of an indicator variable. The power of the indicator variable approach comes from achieving two objectives in one step: (i) it ensures that all variables are sampled at the same state of evolution for all batches, and (ii) it homogenizes the number of samples taken for each batch (K needs to be as equal as possible for all batches).

3. Batch process alignment

Batch processes commonly have unequal durations, since the recipes for automation (or criterions for manual operation) are based on triggers that rarely depend on time. Disturbances to the materials or to environmental conditions (e.g., temperature of chilled water or cooling air) can introduce changes in the magnitude of the driving forces behind the evolution of the process and hence change the total time it takes to finish a given batch. For these reasons, comparison of batch data using time to is hence rarely adequate. Multiple papers have been presented dealing with this issue [4,7–9] referred to as "batch data alignment" proposing methods ranging from the simple re-sampling procedure against an indicator variable, to the very complex Dynamic Time Warping. Batch alignment is in our experience the most time-consuming step during batch analysis, and is necessary unless the dependence with respect to its evolution is disregarded (steady-state assumption).

When the purpose of a given data analysis exercise is to uncover the effect of a given variable at specific points during the evolution of the process (irrespectively of the time it takes the process to get there) it is then imperative to manipulate the data so that the values of the collected variables are representative of the same points of evolution for all batches. This is the primary objective of batch alignment. Having the same number of samples for all batches is a by-product of alignment and not its primary objective. In fact, in practice there is always some variability in the final state of a batch (or a stage of a batch) that makes it difficult to have exactly the same number of samples for all batches. For example, consider a given process that is executed until a temperature of 90 °C is reached, in addition to having different time durations (due for example to differences in total mass in each batch, or variations in heating medium) it is not unthinkable that there will be some variability of the final temperature across batches. An indicator variable approach [3] using temperature as the indicator variable would be appropriate since the evolution of the process from an execution perspective is indicated by temperature. Now even when all the data is re-sampled (for example) at 0.5 °C intervals; if the variation of the final temperature is ±2 °C centred around 90 °C, it means that some batches will have 4 samples less than the average and some will have 4 samples more than the average. Although the data is properly aligned, the inherent variability in the process will prevent all batches from having the same number of samples.

Pharmaceutical batch processes are commonly operated under a specific recipe with known automation triggers for the operation (e.g., "Heat until temperature is 45 °C"). Using an indicator variable is in our experience the best choice for its simplicity due to the natural way the indicator variables can be chosen according to the execution recipe. In practice, each "batch" will be often executed in multiple stages using multiple triggers to automate the operation of the different stages of the batch. The expected value of these triggers is known a priori and they will usually correspond to variables that are expected to have a monotonic behaviour during that stage of operation. These two characteristics (prior known value and monotonic behaviour) make these variables an excellent choice to align against. Furthermore, the operation personnel is familiar with these trigger variables and building a monitoring scheme around these triggers increases the acceptance of the method at the plant floor. Finally, due to the regulatory nature of pharmaceutical processes, these trigger variables are likely part of the design space for a given process and its values will likely be bounded by regulatory approvals; having a monitoring scheme that works around these trigger variables also increases the strength of the argument to use the MSPC framework for design-space monitoring and verification.

The following sections discuss examples where different alignment techniques have been used. And although differences in performance are discussed, the main advantage of the indicator variables is not necessarily due to any superior performance, but more related to reasons given before (ease and acceptance).

4. Case study #1

The manufacture of the active pharmaceutical ingredient (API) can include a complex sequence of reactions and separations. This first case involves a reaction and a distillation of an intermediate pharmaceutical product. The reaction is executed in 8 stages, with 9 variables sampled during the batch. Each lot of reacted material is then transferred to a distillation step. The distillation is executed in 4 phases with 7 variables measured during the batch. The multiple variables sampled for the process have a much different sampling rate. The complete set consists of 65 batches.

It was decided to work as closely with the raw data as possible, since the data had already undergone the manipulation of the compression algorithm. The compression algorithm in this case was not considered harmful since it was done quite appropriately acting as a low-pass filter, extensive discussions on the effect of data compression can be found elsewhere in literature [10,11]. The challenge is to synchronize batches of unequal time duration, which contain variables of unequal sampling rate. Two approaches were taken and discussed in the following sections.