Proceedings of the 9th International Symposium on
Dynamics and Control of Process Systems (DYCOPS 2010),
Leuven, Belgium, July 5-7, 2010
Mayuresh Kothare, Moses Tade, Alain Vande Wouwer, Ilse Smets (Eds.)

Experiences in Batch Trajectory Alignment for Pharmaceutical Process Improvement through Multivariate Latent Variable Modelling

Salvador García-Muñoz*#. Mark Polizzi*.
Andrew Prpich*. Cathal Strain. Adam Lalonde**. Vilmary Negron***.**

**Pfizer Global Research & Development, Groton, CT, 06340, USA*
 # Tel: 860-715-05-78; e-mail:sal.garcia@ pfizer.com.

***Pfizer Global Manufacturing, Groton, CT. 06340, USA*
****Pfizer Global Manufacturing, Barceloneta, Puerto Rico.*

Abstract: In multivariate analysis of batch data, the step known as trajectory alignment (or synchronization) is not solely intended to homogenize the number of samples across batch data. Its primary objective is to standardize the data according to the evolution of the process, irrespective of the number of 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 (multi-rate data). Since trajectory alignment is not necessarily trivial, a simple approach based on the covariance matrix of the scores from a variable-wise unfolded data set is used 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 since the resulting alignment scheme can be performed in real-time for monitoring applications.

Keywords: Batch Process, Multivariate Monitoring, multi-way PCA, multi-way PLS, Alignment.

1. INTRODUCTION

The application of multivariate latent variable models to analyze batch processes has been widely studied and discussed (Garcia-Munoz 2004;Garcia-Munoz et al. 2006;Garcia-Munoz et al. 2008;Gurden et al. 2001;Kourti, Nomikos, & MacGregor 1995;MacGregor & Nomikos 1992;Nomikos 1995;Nomikos 1996;Nomikos & MacGregor 1994;Nomikos & MacGregor 1995a;Nomikos & MacGregor 1995b;Ramaker et al. 2002;van Sprang et al. 2002;Westerhuis, Gurden, & Smilde 2000;Westerhuis, Kourti, & MacGregor 1998;Wold et al. 1998). These techniques have been successfully implemented in industrial settings, with some applications available in the public literature (Chiang & Colegrove 2007;Garcia-Munoz et al. 2003;Neogi & Schlags 1998).

In spite of the maturity of these techniques, there are still misconceptions about the capabilities of the method, and the expectations that batch alignment techniques (Kassidas, MacGregor, & Taylor 1998;Westerhuis et al. 1999) solve the problem of the uneven number of samples per batch. In fact, the problem of having uneven samples across batches has an additional degree of difficulty: the uneven number of samples across variables within the same batch, referred to as multi-rate data (Lakshminarayanan et al. 1996); this problem needs additional attention beyond simple alignment.

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 simple method to assess the potential impact of the dynamics of a process onto the final product quality.

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 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

Copyright held by the International Federation of Automatic Control

MacGregor et al (MacGregor & Nomikos 1992)- it is incorrect to state that the problem is the PCA method. The real problem is the arrangement of the data.

Given that the data is properly arranged in a 2D matrix as described in Nomikos and Macgregor (1994); it has been proven that a PCA model is equivalent to a multivariate time series that implicitly captures the order of the dynamics in the system with time dependent parameters(Garcia-Munoz, Kourti, & MacGregor 2004). This approach also provides a forecasting model with an explicit mechanism for the parameters to adapt to new samples; providing an accurate prediction of the expected future samples (Garcia-Munoz, Kourti, & MacGregor 2004). These forecasting mechanisms are in fact the only way to model batch data with multivariate methods, for process control, design or optimization (Flores-Cerrillo & MacGregor 2002;Flores-Cerrillo & MacGregor 2004;Flores-Cerrillo & MacGregor 2005;Garcia-Munoz, MacGregor, Kourti, Apruzzece, & Champagne 2006;Garcia-Munoz, MacGregor, Neogi, Latshaw, & Metha 2008).

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. This correspondence is discussed in early papers by Nomikos et al, and a simple procedure was proposed 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 is because it achieves 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).

The following section deals with decoupling these two objectives to address the common misconception that batch alignment only refers to the second objective (homogenizing the number of samples).

3. BATCH PROCESS ALIGNMENT

Commonly batch processes 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. Comparison of batch data using time stamps is hence rarely adequate.

If the purpose of a given data analysis technique 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 the same number of samples.

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 (Nomikos and Mac Gregor, 1994) 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 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.

Batch alignment is perhaps the most time-consuming step during batch analysis, and is necessary unless the dependence with respect to its evolution is disregarded (steady-state assumption). Multiple papers have been presented dealing with this topic (Kassidas, MacGregor, & Taylor 1998;Westerhuis, Kourti, Kassidas, Taylor, & MacGregor 1999) ranging from the simple re-sampling procedure against an indicator variable, to the very complex Dynamic Time Warping.

Pharmaceutical batch processes are operated under a specific recipe with known automation triggers for the operation, the approach of using an indicator variable is in our experience the best as is illustrated in the following examples.

4. CASE STUDY #1

The manufacture of the active pharmaceutical ingredient (API) can include a complex sequence of reactions and separations. The 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 stages 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.

The multi-rate nature of the problem commonly arises due to the existence of a compression mechanism in the historian. These compression algorithms (e.g. the Boxcar algorithm) are implemented to reduce the amount of hard drive used to archive the data. Although the presence of this compression layer will limit the frequency of the signal in the data, it can potentially aid the analysis since it will wash out variability that is considered negligible by the operator.

Copyright held by the International Federation of Automatic Control

The criterion that a compression algorithm uses to determine whether to store a sample or not for a given tag is usually based on the change in magnitude being greater than a threshold, or a maximum dead-time (in which a sample will be stored even if it has not changed after a maximum amount of time).

If a certain variable has very few archived points, it might be undesirable to interpolate intensively (to match the number of samples of another one) since no additional information is actually being included into the model.

As a first step in the analysis of this particular application 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 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.

4.1 First alignment approach entirely based on total number of samples for a multi-rate scenario

In order to preserve as closely as possible the number of samples, and the sampling frequency in each stage, the following procedure was followed:

1. Histograms of total number of samples for each variable were made; a median number of samples was calculated for each variable (let s_j be the median number of total samples for variable j).

2. An analysis was made of the median percentage of the total time invested in each of the n stages for each process (8 stages for the reaction, 4 for the distillation). Let f_n be median fraction of total time spent in stage n.

3. For each variable we resample $round(s_j/f_n)$ times during each of the n stages for the process

This re-sampling mechanism yields equal number of samples for all batches, and also some degree of alignment. As it will be shown later, this alignment is much improved when an indicator variable is used.

For the case of the reaction the median fractional time spent for each of the 8 steps were 0.13, 0.04, 0.23, 0.3, 0.03, 0.04, 0.04, and 0.19 respectively. The numbers of samples for each variable for the reaction are listed in Table 1. The median fractional times for the distillation step were 0.32, 0.10, 0.55 and 0.025 respectively; the total number of samples for each variable is given in Table 2.

The raw trajectories for variable TIC-03R10 and the obtained alignment using this first procedure are illustrated in Figs. 1 and 2. An additional trajectory for the used-time at the reaction and distillation was also included in the analysis as well as a block of data with initial conditions and key properties of the incoming material (Garcia-Munoz, Kourti, MacGregor, Matheos, & Murphy 2003).

Table 1. Median number of samples per variable for the reaction step

Tag name	Description	Samples
TIC-01R13/14	Reactor Temperature	85
TIC-03R13/14	Jacket Inlet Temp.	1000
TIT-01R13/14	Upper Reactor Temp.	70
TIT-02R13/14	Jacket Outlet Temp.	450
TIT-06R13/14	Lower Reactor Temp.	70
PIC-01R13/14	Reactor Pressure	65
PIC-02R13/14	Discharge Pump Press.	65
SIC-01R13/14	Agitator Speed	900
WIT-01R13/14	Reactor Weight	970

Table 2. Median number of samples per variable for the distillation step

Tag name	Description	Samples
TIC-01R10	Reactor Temp.	151
TIC-03R10	Jacket In Temp.	2001
TIT-02R10	Jacket Out Temp.	1001
TIT-11R10	Dist-Temp	71
PIC-01R10	Reactor Press	31
WIT-01R10	Reactor Weight	1201
WIT-01L15	Rec. Reactor Weight	621

For the final mode, a multi-block PLS approach is taken due to the severe disparity of samples across variables (each block of data for one variable for all I batches is considered a block). All variables from distillation and reaction are analyzed in a single model due to the strictly sequential execution of these two operations. The model contains 21 blocks and is able to capture 80% of the variability of final yield of the process using 3 principal components.

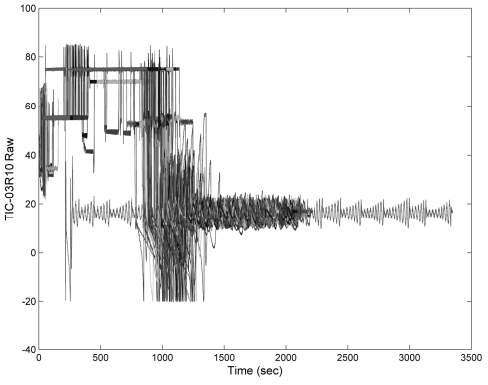

Fig1. Raw trajectories for TIC-03R10

For this model it was noted that the later 2 latent variables captured the most variability in yield. From the analysis of the loadings of these components (not shown) it was clear that the pressure in the reactor, the initial conditions, the properties of the incoming material (TU3), and the agitator speed were strongly related with both components and hence related with the yield of the process.

Copyright held by the International Federation of Automatic Control

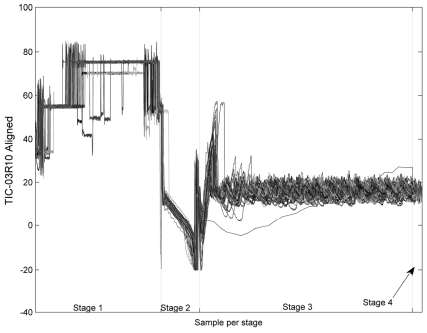

Fig 2. Aligned trajectories for TIC-03R10 using total number of samples procedure

4.2 Second alignment for process monitoring

The results of the previous modelling approach were useful to develop possible strategies to improve the yield of the process; however this model is not useful for monitoring since the total time per batch is unknown at the start of the run.

What is known are the multiple automation triggers used in the execution system (in this case a Delta V® DCS). The trigger variables will surely be measured and in a process improvement case, they represent the desired state of the process hence making it sensible to monitor around them. For this specific example, only the alignment of the distillation step is discussed.

The distillation step is executed in 4 stages. Stage 1 is the primary distillation and the end of it is determined by a certain amount of material extracted and transferred to collection bin. Between stages 1 and 2 there is an addition of a finite amount of additional material (stage 1b). Stage 2 is the cooling down of the system until a thermocouple records -17 C. Stage 3 operated by a finite amount of runtime under slow agitation and stage 4 is the discharge of the unit. Four indicator variables are hence selected from this process: Weight of collecting bin for the first stage, weight of the distillation vessel for stage 1b, temperature for stage 2, and run time for stage 3. Stage 4 is neglected.

A similar analysis as in section 4.1 was carried out with respect to the number of samples taken per variable per stage. This was done in order to determine the total amount of samples to consider per variable per stage. Variables are then re-sampled with respect to the appropriate indicator variable for each of the stages. The aligned profile of TIC-03R10 is shown in Fig 3; note the improved alignment of the data when compared with Fig. 2.

Once the data was aligned, a 5 component model captures ~85% of the total variability in the data. The interpretation of the loadings for this second model is consistent with the expected driving forces acting upon the process. For example, thermodynamic relationships for the primary distillation and heat transfer for stage 1b and 2 are represented in the 1st PC (interpreted so by the relationships between jacket and product temperature, and process pressure).

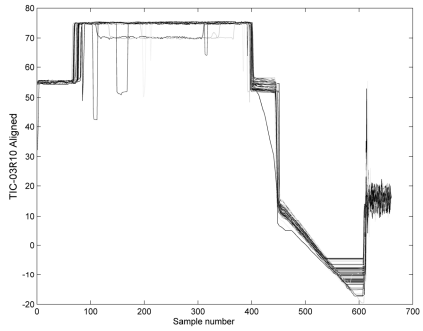

Fig 3. Aligned trajectories for TIC-03R10 using indicator variables

This alignment process can be used for monitoring since all information for alignment is known a priori and real-time alignment is feasible. Fig.5 illustrates an example of the use of this model in a multivariate statistical process control chart (MSPC) (Fig 4 top) where a test batch suffers a deviation from the expected trajectory, the system correctly identifies the instantaneous contributions (Fig 4 bottom). The trajectories of the deviation can be seen in raw numbers in Fig 5.

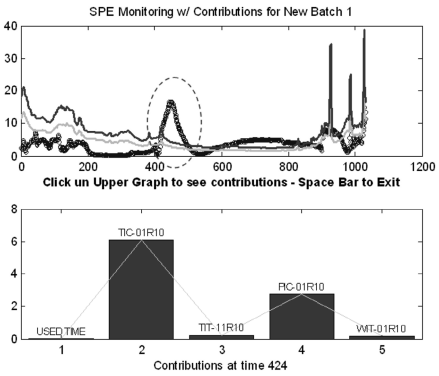

Fig. 4. MSPC chart (top) and instantaneous contributions to time 424 (bottom)

5. CASE STUDY #2

Trajectory alignment is trivial in the involved mathematics (simple re-sampling and interpolating) but can become complex to implement and automate. Intensive bookkeeping of amount of samples and process markers makes batch alignment an exercise that can take significant effort; and there will be cases where the practitioner needs a quick way to assess if this effort is worth while.

This second case study illustrates a simple approach taken to make this assessment during process improvement efforts for the film coating step involved in the manufacture of controlled-release tablets.

The film coating step of oral dosage manufacture is a well understood process driven mainly by the thermodynamics of the solution being atomized and sprayed on the tablets (am

Copyright held by the International Federation of Automatic Control

Ende & Berchielli 2005), and by the mixing process in the tablet bed as the tablets are tumbled in the rotational pan.

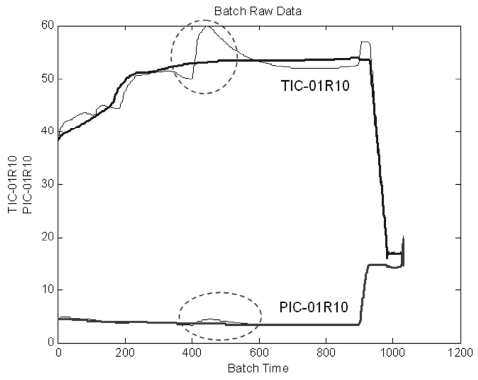

Fig. 5. Raw variables showing the deviations in the two variables detected as irregular

A process improvement exercise required the analysis of historical data for 30 batches. There are 5 process variables that are measured during the run: RH, temperature, and volumetric flow for the inlet air, solution spray rate and exhaust air temperature. The process has a PI controller over the exhaust temperature by manipulating the inlet air temperature. The final product for each of the batches is characterized in the laboratory and labelled according to its performance. Four categories were defined: *Excellent*, *Good*, *Basic* and *Regular* in decreasing order of its desirability.

The film coating step can be said to be a semi-batch process. From the perspective of a product, the process is purely a batch one since the amount of coat in the tablet keeps increasing with time. However, from the perspective of the drying air, the process is known to reach a pseudo-steady state if the interactions of the tablets are neglected (am Ende & Berchielli 2005).

Due to this dual mode in which the process can be analyzed, it was not clear if the analysis of the dynamics of the process was absolutely necessary. The method followed to do this assessment is described next

5.1 Screening method to determine the need to analyze the dynamics of a process

1) All variables within a batch are interpolated to the same number, and all batches are unfolded variable-wise (as suggested by Wold et al (1998)) and arranged in a long vertical matrix of $(K_1+K_2+...+K_i) \times J$. K_i corresponds to the number of samples for i^{th} batch since at this point each batch has a different amount of samples. A PCA model was fitted to this data.

2) The score matrix obtained by the PCA fitted in step 1 has $\sum K_i$ rows and A columns. This matrix is segmented by batch (for each batch takes the corresponding rows of the score matrix and all columns). Metrics of position and spread are calculated for each of the I segments of the score matrix; specifically the mean value per column and the complete variance-covariance (which is re-arranged in a vector of values). The vector of metrics from the segment of the score matrix is then augmented with the mean SPE distance per batch, and mean Hotelling's T^2. For this case, a vector of 7 descriptors was available for each of the I batches, since the PCA was found to have two significant principal components.

3) Perform a PLS model between the matrix of descriptors obtained in 2 and the product quality matrix. Assuming the model captures the necessary patterns in the responses, proceed to step 4. Otherwise stop, the signature in the process is not related to the response.

4) The loadings of the PLS model fitted in (3) are then interpreted, assessing the importance of the variance-covariance of the scores from the PCA model fitted in (2) relative to the importance of the mean values of the scores from the PCA model fitted in (2). If the importance of the variance-covariace of the PCA scores is high, this implies the dynamics of the process do influence the response, otherwise (if all the importance is in the mean values) the dynamics don't matter and only the great mean values of the process parameters are influencing the response.

Fig. 6 illustrates the difference in the behaviour of the scores from the PCA fitted in (2) between batches that resulted in *Excellent* product, and batches that resulted in *Regular* product. It is clear from this plot that the spread of the score values is likely to be related to the difference observed in product performance. The scores for this PCA model will be referred to as $PCAt_1$ and $PCAt_2$ to avoid confusion with the scores from the PLS model fitted in (3).

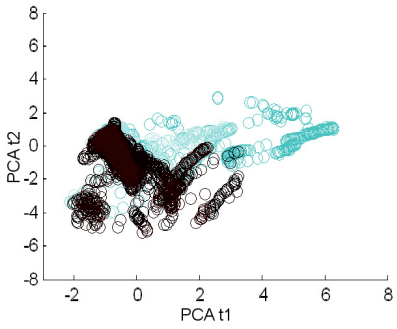

Fig.6 Overlay score plot from PCA model. Black markers are *Excellent* batches, gray markers are *Regular* batches

The loadings for this PCA model (Fig.7) were helpful in understanding the variability in the data around the pseudo-steady state the process operates. The 1st component was interpreted as the behaviour of the controller that would decrease the inlet air temperature (increasing its RH) and flow when the exhaust temperature was rising (event that corresponded with a decrease in spray rate). The 2nd component was interpreted as the thermodynamic relationships between the spray rate, the inlet air temperature and the exhaust temperature (higher spray or lower inlet temp results in lower exhaust temperature).

Copyright held by the International Federation of Automatic Control

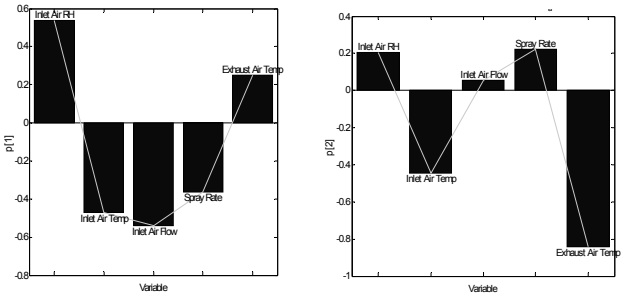

Fig 7. Loadings from the PCA model fitted in (2)

A VIP coefficient was constructed with the loadings from the PLS fitted in (3). This assessment indicated that the most important variables in the model were the variance in $PCAt_1$, the mean value and the variance in $PCAt_2$. Most important of all, it was noted that a t_1-u_1 plot did exhibit a somehow coherent clustering of the batches according to their quality (Fig. 8). These results were indicative that the dynamics of the process (specifically the effect of the disturbances affecting the temperature control loop) were important in the definition of the final quality of the product. A multi-way model was then fitted to fully account for the dynamics in the process.

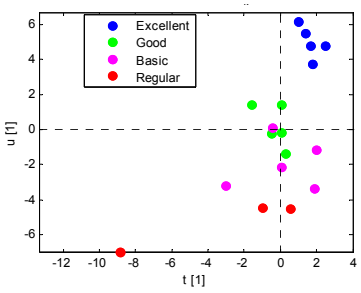

Fig.8 t_1-u_1 score scatter for PLS model fitted in (2)

5.2 Data alignment and multi-way model

The batch data was aligned using the total amount of coating material to be applied as an indicator variable. This is a quantity that is calculated prior to the run and provides an index to perform real-time alignment (which will be of use when and if the model is used for monitoring purposes). Once the data is re-sampled with respect to the amount of coating material applied (which is easily calculated as the integral of the solution spray rate), a multi-way PLS model was fitted between the batch data and the performance of the final product.

This new model provided additional understanding of the process, and captured 90% of the variability in the product performance (see Fig 9). The clustering of batches according to their performance is much better in the score space (Fig 10) when compared to that obtained from the previous PLS model. The last improvement in the score scatters is due to the improved predictability in the data, achieved by a better synchronization of the trajectories.

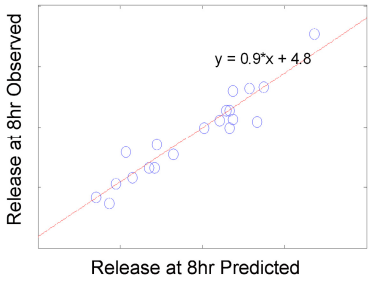

Fig 9. Predicted vs Observed dissolution at 8hr (magnitudes are blinded)

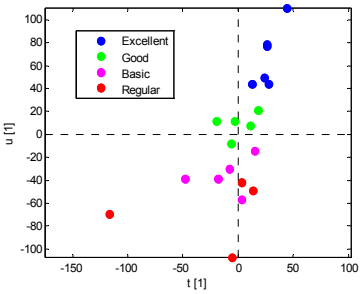

Fig 10. t_1-u_1 score scatter for multi-way PLS model

6. CONCLUSIONS

Batch process analysis is not a trivial exercise not because the methods or the techniques need further development (to make it easier!) but because the dynamics of the process add increased complexity to the system to be analyzed (when compared to the steady state).

In our experience analyzing batch data from pharmaceutical processes, the use of the variables associated with the automation triggers used in the execution recipe as indicator variables is a simple and powerful method to align the batch data. The use of these key variables as indicators of process evolution results in an alignment strategy that can be applied in real-time if the model is to be used for monitoring. For multi-stage processes the analysis will likely require the use of more than one indicator variable.

Two cases are presented where different approaches were used to handle the dynamics in the process, indicator variable alignment proved to be a better approach than time interpolations. A simple procedure was presented to triage the need to align the batch data. This method was applied to the analysis of data from a film coating step and shown to provide an early assessment of the importance of the dynamics. This crude approach needs no alignment of the data and was able to provide an acceptable classification of the batches according to the performance of the final product. The use of an indicator variable proved to be much superior in extracting the dynamic features of the data and hence resulted in a better prediction and classification of the batches.

Overall, we believe that analyzing data from a batch process and performing a proper alignment of the variable trajectories provides detailed process understanding which is key to the assurance of quality in our products.

Copyright held by the International Federation of Automatic Control