Mears, L., Nœrregård, R., Stocks, S. M., Albaek, M. O., Sin, G., Gernaey, K., Villez, K. (2015). Multivariate analysis of industrial scale fermentation data. 12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering (PSE2015/ESCAPE25), Copenhagen, Denmark, May 31 - June 4, 2015, 1667-1672.

Krist V. Gernaey, Jakob K. Huusom and Rafiqul Gani (Eds.), 12th International Symposium on Process Systems Engineering and 25th European Symposium on Computer Aided Process Engineering.
31 May – 4 June 2015, Copenhagen, Denmark

Multivariate Analysis of Industrial Scale Fermentation Data

Lisa Mears[a], Rasmus Nørregård[a], Stuart M. Stocks[b], Mads O. Albaek[b] , Gürkan Sin[a], Krist V. Gernaey[a], Kris Villez[c*]

[a]*Department of Chemical and Biochemical Engineering, Technical University of Denmark, Lyngby, 2800, Denmark*
[b]*Novozymes A/S, Pilot plant, Krogshoejvej 36, Bagsværd, 2880, Denmark*
[c] *Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland*
**kris.villez@eawag.ch*

Abstract

Multivariate analysis allows process understanding to be gained from the vast and complex datasets recorded from fermentation processes, however the application of such techniques to this field can be limited by the data pre-processing requirements and data handling. In this work many iterations of multivariate modelling were carried out using different data pre-processing and scaling methods in order to extract the trends from the industrial data set, obtained from a production process operating in Novozymes A/S. This data set poses challenges for data analysis, combining both online and offline variables, different data sampling intervals, and noise in the measurements, as well as different batch lengths. By applying unfold principal component regression (UPCR) and unfold partial least squares (UPLS) regression algorithms, the product concentration could be predicted for 30 production batches, with an average prediction error of 7.6%. A methodology is proposed for applying multivariate analysis to industrial scale batch process data.

Keywords: Multivariate Data Analysis, Bioprocess, Process Optimisation

1. Introduction

There is a vast amount of batch process data generated in industry, which can be investigated with the aim of identifying desirable process operating conditions, and therefore areas of focus for optimising the process operation. Although multivariate methods are highly established for analysis of large data sets, their application to batch processes is less common due to the additional challenges associated with data dimensionality, as well as high measurement noise. Data mining of such a complex data set requires additional pre-processing stages, however the importance of these steps prior to modelling is often underappreciated (Gurden et al. 2001).

Nomikos and Macgregor (1994, 1995) pioneered the application of multivariate statistical methods to batch processes. With a lack of mechanistic models which can define the non-linear process dynamics and unsteady-state operation, and a lack of sensors which measure key process variables, empirical modelling shows promise for characterizing batch operations (Nomikos and MacGregor 1995).

Fermentation processes are highly sensitive to operational changes, however the underlying mechanisms are often poorly understood. Multivariate methods are therefore suitable tools for gaining process insight (Formenti et al. 2014). Such biological processes are also sensitive to system variations between batches and are often hard to implement in a reproducible manner. In addition, there is evidence to suggest that not only the main fermentation conditions, but also the seed fermentation conditions are vital in order to dictate the resulting yield of the process (Ignova et al. 1999). An additional benefit of multivariate methods is that offline data can be incorporated into the model which helps to account for differences observed between batches (Kourti et al. 1995).

In this work, a 30 batch dataset from a filamentous fungus production process operating at Novozymes A/S is analysed by multivariate analysis with the aim of predicting the final product concentration, which is measured offline at the end of each batch. By investigating the contribution of each variable to the model prediction (Kourti et al. 1995), differences between batches can be diagnosed. This diagnosis procedure can be used to understand the process, and therefore guide future process optimisation efforts.

2. Methodology for applying multivariate analysis to batch data

The multivariate methods require a two dimensional dataset, however batch production processes have dimensions of time, batch number and variable. It is also a requirement for UPCR and UPLS that the data matrix for each batch should have the same dimensionality, however this is not the case due to different data logging and data compression for each variable, as well as different batch durations. These issues must be addressed as part of the data pre-processing. There are then considerations for data scaling, so that all variables have the same weighting in the resulting model. Finally the method of model validation should be discussed in order to validate the conclusions drawn from the regression model. Each of these challenges will now be approached with a discussion of the options available. Figure 1 provides the general methodology for applying the methods discussed.

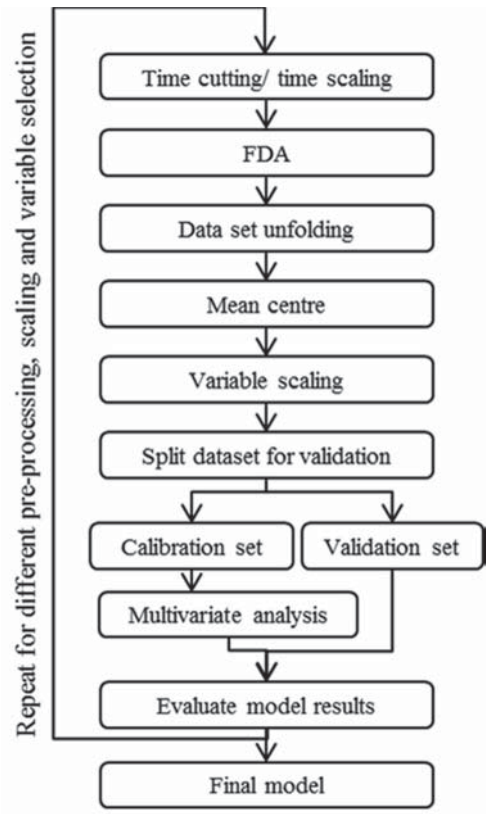

Figure 1: Proposed methodology for multivariate analysis of batch process data.

2.1. Time Scaling Methods

Since batch operations have different durations, the time must be scaled in order to compare multiple batches. The most simple method is to cut the data to the length of the shortest batch, however this is only applicable if the batches show only small variations in duration, otherwise key

data may be lost. There is also the option to cut the data from the beginning, so that the data which is removed is only lag phase data which contains limited relevant dynamics, however the implications of the lag phase can then not be captured, and this may also have an impact on the process.

Alternatively there is the option to linearly scale the time, so that the index of a time unit is scaled such that all batches have the same dimensionality, but with different absolute times. This maintains all data, however it affects the relevance of certain time dependent variables, such as rates or cumulative variables, since the scaling removes this time dependence. Time scaling is highly compatible with functional data analysis (FDA) pre-processing, by fitting the same number of functions across the different batch lengths, which is the method applied in figure 2, and discussed in the next section.

More advanced methods include dynamic time warping which is an example of time series alignment (Keogh and Ratanamahatana 2002). This is a non-linear time scaling, whereby the time series is warped in order to align events in the series, which can be defined by a numerical property of the data, for example a maxima in a variable. This allows batch profiles to be compared relative to key events. As an example, aligning DO limitation profiles in the initial stages, which represents the initial exponential growth phase ending. This method is suitable to aligning events on batch data which can show between batch variations in timing (Kassidas et al. 1998). This method was not applied in this work.

2.2. Data set pre-treatment

It is common that the data logging algorithms result in very different data dimensionality for each variable. In order to avoid large-dimensional datasets, one approach is to log values only after a specified magnitude of change and in addition after a set amount of time. This causes each variable to have a different number of samples, and at different batch times. In order to apply uPCR and uPLS algorithms, data pre-processing stages must create a data matrix for each variable, and each batch with an even time distribution. In addition, variables can be measured with varying degrees of precision, meaning measurements capture different levels of noise. In the literature

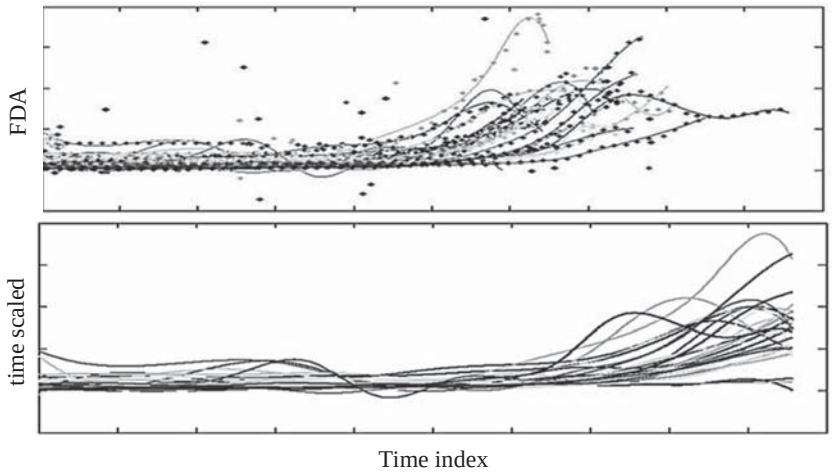

Figure 2: Variable pretreatment, showing the results for FDA (top) plotted with the raw data, and the results of time scaling using functional data analysis (bottom). No axis for confidentiality.

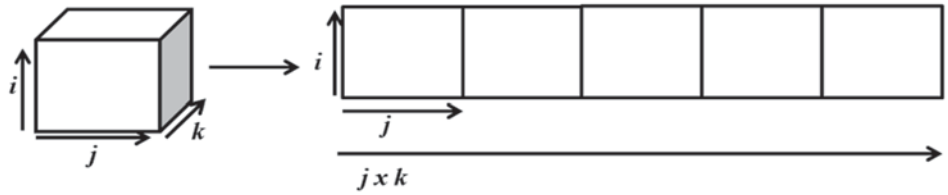

Figure 3: Batchwise unfolding, for i batches, j time, and k variables.

there are different methods used to approach this issue ranging from manual outlier removal (Albert and Kinley 2001), moving average filters, and data interpolation (Le et al. 2012), showing that there is no consensus on the optimal approach.

In this work, the application of functional data analysis (FDA) prior to multivariate methods (Baert et al. 2012) has proven highly valuable in order to greatly reduce the dimensionality of the dataset whilst also filtering noise from the measurements and dealing with outliers in the data in a single processing step. FDA involves fitting of a function across intervals of the data. This method is highly suited, as it allows flexibility in the function output so that the same method can be applied in series to multiple variables. Since this work focusses on non-periodic data, only the spline basis system is considered. Optimal fitting parameters were determined by means of the leave-one-out procedure discussed in James and Silverman (2005).

2.3. Data unfolding

The uPCR and uPLS algorithms apply to a two dimensional data set, therefore dataset unfolding is required (Nomikos and MacGregor 1995). Given a three dimensional matrix with i batches, j time index, and k variables, as shown in Figure 3, unfolding refers to taking slabs from the dataset to create multiple two dimensional matrices of [i x j]. These are aligned to give a two dimensional matrix [i x jk]. This method is most applicable to regression methods for end of batch quality variables, as is discussed in this work.

2.4. Variable Scaling Methods

Column scaling is applied to each time index, so that each time point has equal variance. This is commonly used, however there is the risk that noise is amplified in variables where the overall trend in the data is important but there is high background noise present (Gurden et al. 2001). For example for a cumulative flow profile, implementing column scaling may amplify noise at the start of the batch, where the variance in the values is lower, and lose important information on the final cumulative flow profile as the variance is scaled equally at each time index.

In single-slab scaling all time points for a single variable are scaled for equal cumulative variance. This means that the trajectory shape is maintained more than in column scaling. Figure 4 shows the effect of the variable scaling method on the data profiles.

Model validation methods

In order to assess the resulting model, it is important that suitable model validation is in place, where there is a balance between the fit of the model to the validation dataset, compared to the number of components in the multivariate model. A simple but

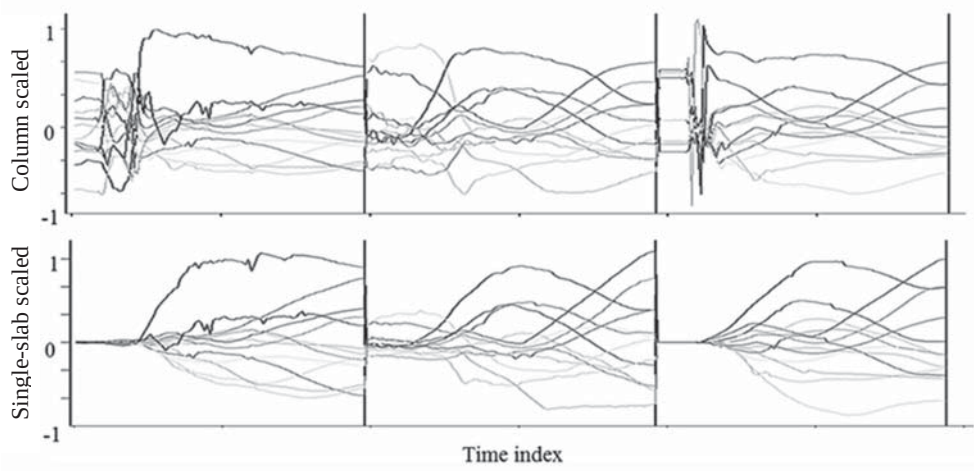

Figure 4: Column scaling (top) and single-slab scaling (bottom) applied to three variables unfolded batchwise. Column scaling affects the cumulative profile, and amplifies noise in the initial time indexes. Variable names and time indexes are excluded for confidentiality reasons.

computationally expensive validation method is leave-one-out, whereby every batch is used as a validation batch. Alternatively the full data set can be split into calibration and validation sets, however the results can then be dependent on the division of the dataset. Leave-one-out validation is suitable for regression applications, since it is then possible to see the effect of individual batches on the model prediction at each calibration stage. This means that if individual batches are outlying, this is seen in the results.

3. Multivariate analysis results from Novozymes A/S industrial dataset

The 30 batch dataset has been analysed by utilising PCR and PLS algorithms implemented in Matlab at Eawag. The discussed data pre-processing stages have been applied, in order to assess the effect on the resulting regression. Many iterations of the modelling were completed, following the methodology in Figure 1, and the effect of the different scaling methods is apparent, as shown by an example in figure 5.

For the final model, time cutting to the shortest batch length was always found to be most effective. Then FDA was applied, with fitting using a roughness penalty (James and Silverman 2005). After dataset unfolding, mean centring was applied followed by

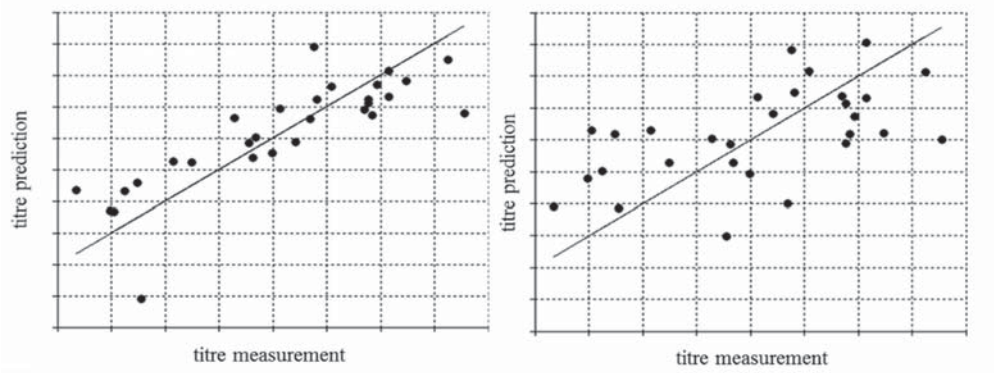

Figure 5: UPCR titre regression models for 30 batch production scale data set at Novozymes A/S. Single slab scaling (left), Column scaling (right). Axis values removed for confidentiality.

single-slab scaling. This resulted in the regression shown to the left in Figure 5.

Based on analysis of the regression coefficients, variables were removed if they had low influence on the resulting model. The accuracy of the model was greatly improved by removal of some variables, which shows that certain variables, although showing high variability, did not show variance relevant to the product concentration prediction, and this therefore affects the model results. The final model included only five online variables, making the model interpretation more straight forward for process optimisation. By analysis of the regression coefficients, it is possible to identify the trend in the data responsible for a higher product concentration prediction. This can lead to identification of potential process optimisation leads.

4. Conclusions

This work discusses how pre-processing is an important and integral stage in multivariate analysis of batch process data, and the choice of methods has an effect on the resulting model, which in this case has been proven by analysis of an industrial fermentation data set from Novozymes A/S. This study was initiated following the observation of practical challenges met during application of conventional statistical process control techniques to an industrial data set. Figure 5 shows that with successful pre-processing it is possible to predict the titre from an industrial fermentation process with an average prediction error of 7.6%. The methodology proposed aims to provide a framework for approaching the multivariate analysis of batch process data.

References

Sarolta A., Kinley R. D. 2001. “Multivariate Statistical Monitoring of Batch Processes: An Industrial Case Study of Fermentation Supervision.” Trends in Biotechnology 19(2):53–62

Baert A., Villez K., and Steppe K. 2012. “Functional Unfold Principal Component Analysis for Automatic Plant-Based Stress Detection in Grapevine.” Funct Plant Biol 39(6):519.

Formenti L R, Nørregaard A, Bolic A, Hernandez D Q, Hagemann T, Heins A, Larsson H, Mears L, Mauricio-Inglesias M, Krühne U, Gernaey K V. 2014. “Challenges in Industrial Fermentation Technology Research.” Biotechnology journal 9(6):727–38.

Gurden S. P., Westerhuis J. A., Bro R., and Smilde A. K. 2001. “A Comparison of Multiway Regression and Scaling Methods.” Chemometr and Intell Lab 59(1-2):121–36.

Hastie T., Tibshirani R., Friedman J. 2009. The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition.

Ignova M., Montague G. A., Ward A. C., Glassey J. 1999. “Fermentation Seed Quality Analysis with Self-Organising Neural Networks.” Biotechnology and bioengineering 64(1):82–91.

Ramsey J., and Silverman B. 2005. Functional Data Analysis.

Kassidas A, MacGregor J F, and Taylor P A. 1998. “Synchronization of Batch Trajectories Using Dynamic Time Warping.” AIChE Journal 44(4):864–75.

Keogh E., Ratanamahatana C A. 2002. “Exact Indexing of Dynamic Time Warping.”, Knowledge and Information Systems, 7(3), 358–386.

Kourti T, Nomikos P, MacGregor J F. 1995. “Analysis, Monitoring and Fault Diagnosis of Batch Processes Using Multiblock and Multiway PLS.” Journal of Process Control 5(4):277–84.

Le H, Kabbur S, Pollastrini L, Sun Z, Mills K., Johnson K, Karypis G, Hu W. 2012. “Multivariate Analysis of Cell Culture Bioprocess Data--Lactate Consumption as Process Indicator.” Journal of biotechnology 162(2-3):210–23.

Nomikos P, MacGregor J F. 1994. “Monitoring Batch Processes Using Multiway Principal Component Analysis.” AIChE Journal 40(8):1361–75.

Nomikos P, MacGregor J F. 1995. “Multi-Way Partial Least Squares in Monitoring Batch Processes.” Chemometrics and Intelligent Laboratory Systems 30(1):97–108.