European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

The Integration of Process and Spectroscopic Data for Enhanced Knowledge Extraction in Batch Processes

C. W. L. Wong[a], R. E. A. Escott[b], A. J. Morris[a], E. B. Martin[a*]

[a]Centre for Process Analytics and Control Technology
School of Chemical Engineering and Advanced Materials
University of Newcastle, Newcastle-upon-Tyne, NE1 7RU, UK

[b]GlaxoSmithKline Chemical Development, Tonbridge, TN11 9AN, UK

Abstract

Batch process performance monitoring has been achieved primarily using process measurements with the extracted information being associated with the physical parameters of the process. More recently, there has been an increase in the implementation of process spectroscopic instrumentation in the processing industries. By integrating the process and spectroscopic measurements for multivariate statistical data modelling and analysis, it is conjectured that improved process understanding and fault diagnosis can be achieved. To evaluate this hypothesis, an investigation into combining process and spectral data using multiblock and multiresolution analysis is progressed. The results from the analysis of an experimental dataset demonstrate the improvements achievable in terms of performance monitoring and fault diagnosis.

Keywords: Multiblock; Multiresolution analysis; On-line monitoring; Batch processes.

1. Introduction

Since the introduction of the Process Analytical Technology (PAT) initiative, companies in the processing industries are increasingly aware of the need to attain a detailed understanding of their processes and products. The goal of PAT is to build quality into the process and remove the final step of testing the final product, thereby achieving the ultimate goal of parametric release. To deliver against this objective, an enhanced understanding of the process and the product is required. One approach to realising this objective is through the implementation of on-line spectroscopic analysers. A second aspect is the need for on-line real time process performance monitoring.

Traditionally batch process performance monitoring (Neogi and Schlags, 1998; Martin and Morris, 2002) has been performed primarily using process measurements with the extracted information being associated with the physical parameters and/or inferred chemical parameters of the process. More recently, there has been an increase in the use of process spectroscopic instrumentation for the monitoring of a process (Gurden et al., 2002). Spectroscopy provides real-time, high-quality chemically rich information, but in

* Author to whom correspondence should be addressed: e.b.martin@ncl.ac.uk

most studies, the data is analysed independently of the process data. By integrating the process (physical state) and spectroscopic (chemical state) measurements for multivariate statistical data modelling and analysis, it is hypothesised that improved process understanding and fault diagnosis can be achieved. To evaluate this belief, an investigation into combining the two data forms using multiblock and multiresolution analysis is conducted. The results of the combined analysis are compared with those attained from separate analyses undertaken on the spectral and process data.

A number of approaches to combining Principal Component Analysis (PCA) and Partial Least Squares (PLS) with wavelet analysis have been proposed (Bakshi, 1998). Current methods reported in the literature have involved the selection of an appropriate scale as the basis of the monitoring scheme after the application of the wavelet transformation or alternatively applying the projection method to the decomposed scales, complicating the interpretability of the final process representations. To address this level of complexity, multiblock analysis is considered. Multiblock methods can enhance the identification of the underlying relationships between several conceptually meaningful blocks thereby summarising the relevant information both between and within the blocks in a single representation. To demonstrate the potential of the developed methodology, process and UV-visible data from a batch mini-plant is considered.

2. Methodology for Data Integration

In a typical batch production environment, process information and spectra are normally acquired in separate data historians thus dividing the data into two distinct data blocks. These two blocks are conceptually meaningful since the same object is measured but the description of the state differs. Two approaches are developed in the subsequent sections and compared for combining the process and spectral data.

2.1 Multiblock Method

For the first approach, the spectroscopic and process data are integrated using multiblock analysis, more specifically consensus PCA (CPCA) (Westerhuis et al., 1998). Figure 1 provides a schematic of the proposed integrated on-line monitoring scheme. The process and spectral data are divided into two base blocks and CPCA is applied. More specifically a starting super score t_T is selected as the first column for one of the blocks and this vector is regressed on both blocks to give block variable loadings. The block scores t_b are then calculated and combined into a super block **T**. The super scores are then regressed on the super block to give the super weights of the block scores with the super weight being normalised to unit length and a new super score is calculated. The procedure is repeated until the super score converges. Both the super and block scores are then used for the monitoring of the performance of the process.

2.2 Multiresolution Analysis

For the second approach, integration is performed as per the first approach but the spectral data is first pre-processed using wavelet analysis. Most data generated from chemical processes is inherently multiscale and multivariate in nature. Spectral data is no exception and usually comprises a large number of wavelengths thus the interpretation of such a large and complex data matrix requires advanced techniques to reduce the dimensionality and complexity of the problem. Wavelets have been proven to be a useful tool to denoise signals and extract multiscale components (Trygg and Wold, 1998; Teppola and Minkkinen, 2000).

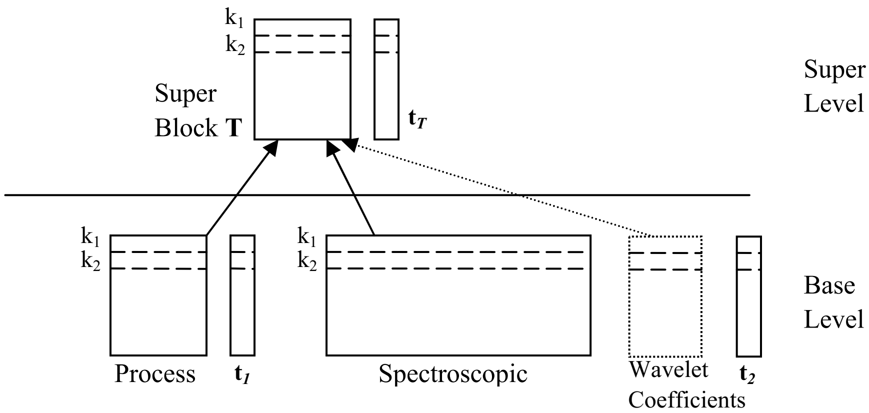


Figure 1. Integrated on-line monitoring scheme by CPCA

In the second approach, the spectral data is decomposed using the discrete wavelet transform with the original signal being recursively decomposed at a resolution differing by a factor of two from the previous step. During the decomposition, the smallest features (noise) are first extracted, resulting in an approximate signal. From this approximation, new features are extracted, resulting in an ever more coarse approximation. This continues until the signal has been approximated to the pre-selected level. The differences are stored as wavelet coefficients. If all wavelet coefficients are used, the original signal can be perfectly reconstructed. In Figure 1, the dotted section is included into the CPCA but not the spectroscopic block. The size of the dataset is significantly reduced however the details are retained with the multiscale components being extracted.

3. On-line Monitoring

3.1 Process Description

A simple reaction of nitrobenzene hydrogenation to aniline is considered. Eight experiments were performed of which six batches formed the nominal data set. Seven process variables were recorded every second including reactor temperature, pressure, agitator, H_2 gas feed, jacket inlet and outlet temperatures and flow rate of heating fluid with the UV-Visible spectra being recorded every 30 seconds. Two batches with pre-defined process deviations were also run. The first of these, batch 7, was discharged with 10% less catalyst to simulate a charging problem and to simulate a series of temperature control problem. The second batch, batch 8, simulates a series of agitator speed and pressure loss problems. The changes reflect both a change to the process as well as to the chemistry. In the application, Daubechies-4 wavelet with five decomposition levels was chosen with the last level of wavelet coefficients being considered as the spectral block as opposed to the original spectra.

3.2 Data Pre-processing

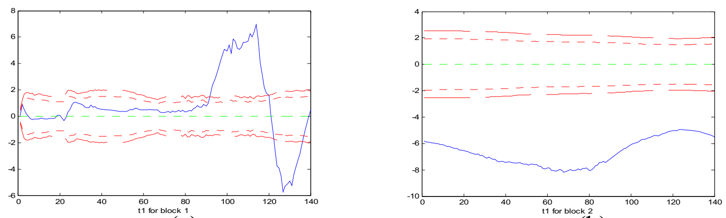
One of the challenges of data integration is to time align the disparate data sets. The process measurements may be recorded with a sampling interval of seconds but the time frame for the spectroscopic measurements is typically larger. In this study to realise the more rapid detection of a fault, a sampling rate of ten seconds was selected, hence interpolation of the spectral data was necessary. Additional pre-processing of the UV-Visible spectra was required since it exhibited a baseline shift therefore a baseline correction was applied to the spectroscopic data.

Since the process and spectral data blocks are three-dimensional matrices, **X** (I x J x K), the first step is to unfold the data to a two-dimensional array. The approach of Nomikos and MacGregor (1994) was adopted resulting in a matrix of order (I x JK). Auto-scaling was then applied to the unfolded matrices for the removal of the mean trajectories. A weighing factor was also introduced at this stage to ensure the variance of each block was unity. The weighting factor to achieve this was $1/n^{1/2}$ where n is the number of variables in a block. The next step was to re-arrange the matrices into a matrix of order (IK x J) to enable the application of the Wold et al. (1998) approach. By adopting this procedure the issue of unequal batch length monitoring and the need to consider how to handle on-line performance monitoring is reduced. CPCA is then applied to the pre-processed data blocks with the principal component scores being used for monitoring.

3.3 Results

3.3.1 Multiblock Approach

The process and spectral block scores for the first principal component for batch 7 are shown in Figure 2a and 2b. The temperature control problem is observed from Figure 2a and verified using the contribution plot (Figure 3). It is expected that a slower reaction would occur when less catalyst is charged into the vessel thereby affecting the overall kinetics of the reaction. This effect is observed in Figure 2b (spectral block) where the trajectory is observed to be out of control throughout the whole process.



(a) (b)

Figure 2. Block scores of principal component one. (a) Process; (b) Spectral

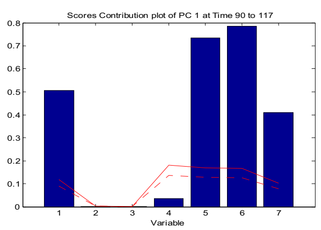


Figure 3. Contribution plot for the process block scores

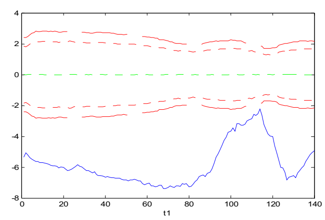


Figure 4. Super scores for principal component one for batch 7

The super scores of principal component one were interrogated. Figure 4 illustrates the advantages of the multi-block approach. It summarises the deviations from the process and spectral blocks. Figure 5 shows the super scores of principal component one for batch 8. This batch has mainly process disturbances as observed from the process block scores (Figure 6) since the spectral block scores (Figure 7) revealed no out-of-control signal. Most of the process disturbances are detected from the super scores however the agitator disturbance during the period 43 – 49 was not detected as the main source of failure (Figure 8). This result will be compared with the multiblock-wavelet approach.

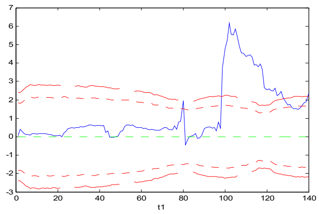


Figure 5. Super scores of principal component one for batch 8

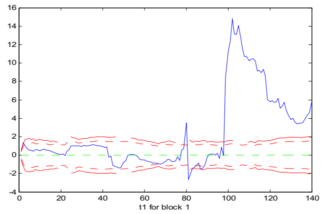


Figure 6. Process block scores of principal component one

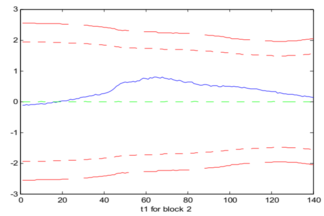


Figure 7. Spectral block scores of principal component one

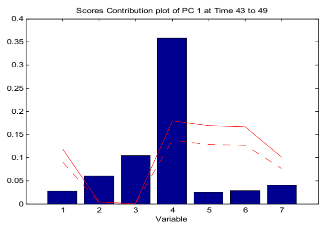


Figure 8. Process block scores of principal component one

3.3.2 Multiblock-Wavelets Approach

For the multiblock-wavelet pre-processing approach, the number of variables (wavelengths) for the spectral data was significantly reduced from the original number of wavelengths, i.e. 216 to 14 wavelet coefficients, resulting in the data being compressed 15-fold. However, the process features are retained as evidenced from the coefficients shown in Figure 9 for batch 8.

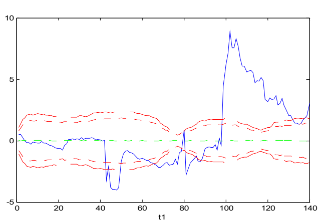


Figure 9. Super scores of principal component one for batch 8

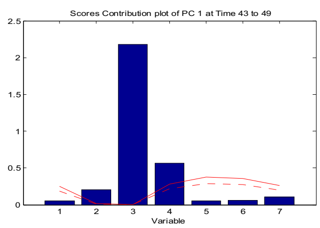


Figure 10. Contribution plot for process block scores

Interrogating the super scores of principal component one, it was observed that a similar result is obtained. The focus is on the agitator disturbance during time period 43 to 49 where it shows an out-of-control signal that was not observed from the multi-block analysis. The contribution plot of the process block scores (Figure 10) confirmed the finding that the fault was primarily due to the failure of the agitator (variable 3) which consequently affected the reactor pressure (variable 2) and H_2 gas feed (variable 4). The approach has been shown to have improved fault detection capability.

4. Discussions and Conclusions

The area of integrated data monitoring has become increasingly more important as increased amounts of data from different data structures are recorded. However, the extraction of information and hence knowledge from such combined data structures is limited. The development of an integrated framework can help in the understanding of the process more than that of an individual model. While further fault diagnosis is required, the integrated model allows tracking back to the base models thus to address the problem accordingly. More specifically in this paper, a successful application to data integration has been proposed where the chemical and physical information are incorporated into the model but interpretation is made simpler in a single representation. Multiblock and wavelet transformation are combined providing a powerful combination of dimensionality reduction and data compression. The correlation between blocks and the multiscale nature of data were also considered. The challenges of time alignment, data scaling and weighing between blocks were discussed.

References

Neogi, D. and C. Schlags, 1998, Multivariate statistical analysis of an emulsion batch process, Industrial & Engineering Chemistry Research, 37, 3971.

Martin, E. B. and A. J. Morris, 2002, Enhanced bio-manufacturing through advanced multivariate statistical technologies, Journal of Biotechnology, 99, 223.

Gurden, S. P., J. A. Westerhuis and A. K. Smilde, 2002, Monitoring of batch processes using spectroscopy, AIChE Journal, 48, 2283.

Bakshi, B. R., 1998, Multiscale PCA with application to multivariate statistical process monitoring, AIChE Journal, 44, 1596.

Westerhuis, J. A., T. Kourti and J. F. MacGregor, 1998, Analysis of multiblock and hierarchical PCA and PLS models, Journal of Chemometrics, 12, 301.

Trygg, J. and S. Wold, 1998, PLS regression on wavelet compressed NIR spectra, Chemometrics and Intelligent Laboratory Systems, 42, 209.

Teppola, P. and P. Minkkinen, 2000, Wavelet-PLS regression models for both exploratory data analysis and process monitoring, Journal of Chemometrics, 14, 383.

Nomikos, P. and J. MacGregor, 1994, Monitoring batch processes using multi-way principal component analysis, AIChE Journal, 40, 1361.

Wold, S., N. Kettaneh, H. Friden and A. Holmberget, 1998, Modelling and diagnostics of batch processes and analogous kinetic experiments, Chemometrics and Intelligent Laboratory Systems, 44, 331.

Acknowledgements

Chris Wong would like to acknowledge the EPSRC, GlaxoSmithKline, the UK ORS Scheme and CPACT for financial support of his PhD.