The introduction of process chemometrics into an industrial pilot plant laboratory

S.P. Gurden *, E.B. Martin, A.J. Morris

Centre for Process Analysis, Chemometrics and Control, Department of Chemical and Process Engineering, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK

Received 30 August 1997; revised 6 March 1998; accepted 9 April 1998

Abstract

Process chemometrics is the application of multivariate statistical methods to industrial process data characterised by a large number of correlated process measurements. In this paper, we aim to show how multivariate techniques have been used in a pilot plant environment with the objective of increasing the general understanding of the process despite having access to limited data. The use of process trajectory plots to follow the operation of the plant are discussed, along with statistical indicators for the detection and diagnosis of process disturbances. The effect of process conditions on product quality is analysed using cross-correlation with latent variables and significant process variables and time delay structures are identified. The experience and process understanding gained by the pilot plant staff has enabled them to propose the installation of new sensors and analysers based upon sound business benefits. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Process chemometrics; Multivariate statistical methods; Industrial pilot plant laboratory

1. Introduction

Modern industrial process plants are capable of monitoring large numbers of process variables at frequent time intervals. Utilised properly, this data can provide a wealth of information leading to a deeper understanding of the process with the ultimate aim of increasing the output of good, consistent quality product, minimising the use of energy and raw materials, reducing effluent discharges and through these increasing business profits. The main problem in the analysis of process data is that the recorded information usually consists of many highly correlated variables. The real process, however, may actually be governed by relatively few underlying factors relating to combinations of certain key process variables. Traditional univariate techniques are unable to utilise information relating to the variable interactions and so multivariate techniques must be applied [1–12]. These methods operate by reducing the dimensionality of the collinear data so as to include only as many factors as are necessary to describe the important variation within the data. Typical objectives of process chemometrics or multivariate statistical process control (MSPC) include process monitoring to ensure overall production control; fault detection and diagnosis; determination of key process variables; and the generation of inference models used to forecast and optimise product quality.
Many scientists and engineers view MSPC methodologies as another, but more advanced technique of statistical process control which can be applied to manufacturing processes, but with limited success. The analytical power of the important underlying basis of MSPC, principal component analysis (PCA)
and partial least squares (PLS) modelling seem to be overlooked, possibly by the including of the term ‘control’ in MSPC. The study described in this paper is based upon the introduction of multivariate statistical analysis techniques into the pilot plant laboratory, looking to optimise the production performance of commercial plants through design of experiments and data analysis. The eventual aim is to ‘map’ the pilot plant operation onto real production plant operations to achieve optimal production.

The paper considers data from an industrial pilot plant which has been designed to be representative of a full-scale production unit. The pilot plant provides access to far more process variables than those available on the full-scale production plant. Indeed, one aim of the analysis is to determine which process variables have an important influence on production and product quality. If this can be achieved it should be possible to make recommendations as to the most important variables for production plant performance, monitoring and control. Further data from a number of other pilot plant runs using different operating regions was available, but are not presented here as the results are merely confirmatory in nature. The specific pilot run presented here contains typical value ranges for all the process and quality variables.

2. The pilot plant data

The data for one day’s operation of the pilot plant is characterised by three stages: (a) a gradual start-up period during which the plant utilities are increased to operating level; (b) a stable period during which normal operating conditions (NOC) are reached and product sampling occurs and (c) a shut-down period during which plant utilities are switched off. A set of 24 process variables relating to a range of conditions across the process are measured every minute over a 12 h time-span, giving a $72 \times 24$ matrix of process variables. Owing to the proprietary nature of the data, these process variables will be referred to by numbering (i.e., PVs 1–24).

Fig. 1 shows the time profiles for the data set. For clarity, the profiles have been plotted on three separate graphs (grouped according to measurement range). It is noted that the process variables differ both in the severity of the noise present and the sensitivity of measurement. Some variables show step increments typical of low resolution measurement and many of the variables exhibit small variability at steady state.

In addition to the process measurements, eight product samples were taken at irregular intervals once the pilot plant operators considered steady state to have been reached. Four measurements relating to aspects of product quality were taken from each sample (QVs 1–4) and are given in Table 1. The relative scarcity of quality data in comparison to process data is a problem common in the analysis of industrial processes and is discussed below. In order to provide appropriate data for subsequent analysis an experimental design was carried out.

3. Process data analysis

3.1. Data pre-processing

The matrix $X$ consists of $72 \times 24$ process variable measurements with each column giving a time profile for a particular process measurement. One common problem with industrial data is that of missing values [13]. Fig. 2 shows some process variables measured between 16:30 and 18:00. A problem with the data collection system, apparently linked to the shut down of PV 1, meant that no values were logged between 17:10 and 17:15. These values need to be filled in before computation on the matrix can be carried out. Data interpolation methods such as smoothing splines [14], time series modelling [15] or record matching using historical data are possible.
However, these were found to be unnecessarily complex in this case. As this particular problem occurred during a relatively uninteresting area in terms plant production (i.e., during shut-down), it was decided to use simple local linear interpolation to fill in the missing values. Another option would have been to remove the offending time region altogether, although this would result in a false disturbance in the concatenated time profiles.

The process variables which make up the data relate to a variety of physical measurement units and some method of scaling must be used in order to avoid giving undue significance to those variables with high absolute values [16–18]. The most common method of achieving this is to autoscale the data by subtracting the mean and dividing by the standard deviation for each column:

\[ ^*X_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \]  

(1)

where \(^*X\) is the \(I \times J\) autoscaled matrix and \(\bar{x}_j\) and \(\sigma_j\) are the mean and standard deviation of the \(j^{th}\) column of the original matrix, \(X\). Thus, each column of the autoscaled matrix has zero mean and unit standard deviation.

3.2. Process trajectory plots

As a method of preliminary data exploration, principal component analysis (PCA) [19,20] was performed on the data. The aim of PCA is to reduce the dimensionality of the problem in order to examine the important trends underlying the multivariate system. The pre-processed matrix, \(^*X\) (\(I \times J\)) is decomposed into a set of scores, \(T (I \times K)\), and loadings \(P (J \times K)\), where \(K\) is the number of principal components, chosen so as to explain the important variation in the data using as few PCs as possible. Almost 80% of the variance in the data is explained by the first two PCs. A scores plot for PC 1 vs. PC 2 is shown in Fig. 3. The process can be seen to follow a trajectory from plant initiation (1), through the gradual start-up phase (2), to the stable sampling...
period (3) and then to plant shut-down (4) and end (5). PC 1 describes the main variation in the process (i.e., horizontal) as the region corresponding to the normal operating conditions (NOC) is attained and then left. During this time, although the values of many process variables are changing, the overall statistical relationship between the values is changing in one direction only. This direction, defined by PC 1, can be interpreted by using the corresponding loadings plot for PC 1 vs. PC 2, given in Fig. 4. The loadings indicate which variables contribute most strongly to each PC. The main influences on PC 1 are the variables in the top right-hand corner (process variables 9–11 and 17–21), which correspond to two specific sections of the pilot plant. The time profiles for these variables are closely correlated and generally follow the same pattern of a gradual increase during the morning until steady operation is reached, then a falling off period during shut-down. Process variables 1, 4, 14 and 15 are seen from the loadings plot to have a strong influence on PC 2. This PC mostly describes variation at the very beginning and end of the process, during which the utilities which drive the plant are switched off and the process is essentially non-functional.

Process variables 2 and 12 show a strong negative correlation with the rest of the process variables. The time profiles given by Fig. 5 show that, indeed, these two variables do not conform to the common pattern. In discussions arising from this analysis, it was revealed that problems had occurred previously with the sensor for PV 2. This variable can be seen to contribute minimal information and is probably best ignored at present. PV 12 represents an unusual process attribute isolated from the main plant operation and so would be expected to act as an outlier on the loadings plot.

As demonstrated above, the off-line (and by extension on-line) graphical output from a PCA on the process data can assist the plant operators in a number of ways. The process trajectory plot allows the plant operator to follow plant progress throughout the day. The use of historical data in conjunction with this plot could be used to alert the operator to any deviation from normal plant operation. Correlations between process measurements are highlighted, indicating relationships which may not have been previously apparent. The close clustering of certain variables may point to a duplication of some process information. In situations where these measurements are very costly to record, the plant operator may decide to reduce the number of sensors being reassured that one measurement is a good indicator of the others. In contrast, by identifying outlying variables the operator is alerted to possible sensor or process malfunctions.

3.3. Determination of normal operating conditions

For the pilot plant data discussed here, it is important to be able to judge when steady state operation is reached and when experimental data monitoring can take place. Representative plant output sampling should only occur during this period of operation. Having performed PCA on the data, two statistical indicators commonly used in process chemometrics, the $T^2$ and $Q$ values, are used [5,8,9]. The $T^2$ statistic gives a measure of variation within the PCA model.
and so gives information as to when an event has statistically anomalous values which do, nevertheless, fall within the defined model space. The Q residual, also known as the squared prediction error (SPE), describes the residual variation and, therefore, how well the PCA model fits the data at each time point. An event with a significantly high Q residual contains information which falls outside of, and so is not recognised by, the defined model space. It is generally useful to consider both of these statistics in concert, as a process disturbance may be signalled by either or both of them. By using standard probability distributions, statistical significance limits can be calculated for automatic event flagging (95% limits are used here) [21].

Fig. 6 shows a dual plot of $T^2$ and Q vs. time. Six PCs, describing 97.07% of the total variance were selected for the PCA model. From time 06:00 to around 11:00 the plant is in start-up mode and a number of process disturbances can be seen. Both statistics indicate that the plant had stabilised by around 12:00 and continued in a state of statistical control until around 16:30 when a shut-down was initiated. Using this information, normal operating conditions were defined as being between 12:00 and 16:30, with these two points being marked as NOC start (S) and finish (F) on the process trajectory plot in Fig. 3. From this figure, it can be seen that the data variation for the normal operating conditions time period occurs within a relatively small region in
comparison to the entire data set and, more signifi-
stantly, constitutes only 37.5% of the overall run time.
In discussions arising from this analysis, the plant
operators and experimental staff have decided to in-
crease the overall run time in order to allow a longer
period for product sampling during stable operating
conditions.

3.4. Analysis of normal operating conditions

Once the normal operation region has been de-
defined, it is appropriate to create a new model for that
region only, since it is here that the pilot plant aims
to represent the full-scale plant and here that repre-
sentative product sampling occurs. Some of the pro-
cess variables have low or zero variance in this re-
region, but were included in the analysis. Although
these variables have little statistical interest within a
single data set, the absolute values may have a sig-
nificant effect on the product quality. This is espe-
cially the case if data sets from different days are to
be compared and so it is important that these process
variables are included. Generally speaking, it is not
always expedient to delete low or zero variance vari-
able, as these may display information relating to a
process fault during subsequent plant operation and
which may otherwise be missed.

Fig. 7 shows a dual plot of $T^2$ and $Q$ vs. time for
the new model, for which six PCs were used. The
process can be seen to be generally in control, with
the exception of a few short disturbances. A useful
way of determining which process variables are re-
sponsible for these disturbances is to use contribu-
tion plots [22–24]. Here, the data is weighted using
the loadings from a PCA, and a bar chart of the pro-
cess variable contributions to a particular PC may be
plotted. The contribution of process variable $i$ to PC
$k$ at time point $j$ is given by:

$$ c_{ijk} = \mathbf{X}_j \mathbf{P}_{ik} $$  \hspace{1cm} (2)

where $\mathbf{P}$ is loadings matrix from a PCA on the
autoscaled data.$^\dagger$ $\mathbf{X}$.

To investigate the cause of the disturbance indicated
by the $T^2$ plot at time 16:18, the contribution
plots for the first few PCs were examined. The con-
tributions to PCs 1 and 2 showed little of significant
interest since they tend to represent average plant be-
aviour. In contrast, the contributions to PC 3, given
in Fig. 8, show process variable 3 to be very signifi-
cant at this time point. Fig. 9 shows the time profile
for this variable around the disturbance time and a
jump can be observed in precisely this region. The
on-line observation by plant operators of this event
enables a decision to be made whether to take action
or not in rectifying the relevant process unit. Distur-
bances in the $Q$ plot may be investigated in a similar
fashion by calculating the residual contributions. The
pilot plant operators now accept that the use of the $T^2$
and $Q$ statistics in conjunction with contributions
plots provides a level of monitoring and fault diag-
nosis which would not be possible by simple visual
inspection of the raw process time series. The meth-
ods described can be routinely used for on-line pilot
plant monitoring and performance evaluation.

4. Production optimisation

An important aim of the process analysis being

\hspace{1cm} Fig. 9. Time profile for PV 3 around the diagnosed disturbance period.

\hspace{1cm} Fig. 8. Process variable contributions to PC 3 at time 16:18.
A quality variable measured at time $i$ will be dependent upon the value of process variable $j$ measured at time $i - n_j$, where $n_j$ is the associated time delay. One method of determining the value of $n_j$ which requires no special plant operation is to use cross-correlation [15, 25], although this assumes that both the process and quality variables are measured with equal frequency and the plant is operating at steady state. The initial correlation coefficient ($r_0$) between the unshifted process and response vectors is first calculated. The response vector, $y$, is then shifted back by one time point in relation to $x$ and the correlation coefficient, $r_1$, calculated. This is continued until a maximum time shift, typically around 10–20% of the total number of data points for this plant, has been completed. Fig. 10 shows a schematic for the cross-correlation analysis using a response vector of

![Cross-correlation schematic](image)

**Fig. 10. Schematic for cross-validation analysis.**

The relationship between the set of quality characteristics, $Y$, and the set of process variable time profiles, $X$, is not instantaneous, but involves a time delay between the effect of the process conditions on the product and the product exiting the system and being analysed.

![Interpolated data for QVs 1 and 3](image)

**Fig. 11. Interpolated data for QVs 1 and 3.**
fixed length against a larger vector of predictor variables, as is the case here. A common problem with applying cross-correlation in this instance is that the quality variables, given in Table 1, are measured relatively infrequently and at irregular intervals. In order to apply cross-correlation between these quality variables and the process variable data, one approach is to expand the measured quality variables into a full time series, using either linear or spline interpolation. Despite being the less sophisticated method, linear interpolation was found to be generally more successful with this particular data set and was used for the experiments described below. Fig. 11 shows the interpolations for QVs 1 and 3.

A subset of \( \mathbf{X} \) was created, to include all events from the time of the first product sample (13:45) up to the last (15:35) (see Table 1). PCA was performed on this data and four PCs retained, explaining 81% of the variance. The scores vs. time plots for the first two PCs are shown in Fig. 12. It is notable that the PC 1 scores show a steady decline with time indicating that the plant is not in steady state operation. This highlights a problem with the daily operation of this pilot plant which must be addressed by longer run times and optimal start-up conditions in order to allow the plant sufficient time to stabilise. For the purposes of determining the relationship between process conditions and product quality, however, this variation in the process conditions is incorporated into the analysis.

Cross-correlation between each of the four scores vectors and the four interpolated quality vectors was then carried out in order to determine any delayed relationships. A maximum time delay of 20 min was
considered appropriate. The cross-correlations for quality measurements 1 and 3 against PCs 1 and 2 are shown in Figs. 13 and 14. QV 1 shows little correlation with PC 1, but a strong correlation with PC 2 with a time delay of approximately 9 min. This may be interpreted by considering the loadings plots for PCs 1 and 2, given in Fig. 15. Process variables which feature strongly on the loadings for PC 2 (i.e., 1, 3, 7 and 18) can be seen to have a delayed influence on QV 1 (and, inversely, on QV 2). QV 3 shows an immediate strong correlation with PC 1 and also a 11 min delayed response to PC 2. Due to the modest amount of quality data used to carry out the analysis, the cross-correlations are only likely to give approximate relationships. However, a definite trend can be established, whereby the process variables contributing to PC 1 (and influencing QVs 3 and 4) are found to be those measured near the end of the process (i.e., downstream), and those contributing to PC 2 (influencing QVs 1 and 2) are those measured upstream. The suggested time scales are corroborated by existing process knowledge of average residence times and transportation speeds within the system. By returning to the pilot plant and applying engineering knowledge to these results, insight is provided into the physical relationships within the chemical process itself. Further work is being carried out in order to confirm and extend upon these results. In particular, it is noted that some process variables may be important and yet be so tightly controlled that they do not change much during the sampling period and appear, therefore, to have minimal effect on output quality. For this reason, recommendations have been made to the plant operators for a series of experimentally designed runs during which process conditions are varied systematically so as to ensure that all process condition/product quality relationships are included.
5. Conclusions

Process chemometrics provides a wide range of multivariate methods which may be used both on-line and off-line to track and extract valuable information in an industrial pilot plant environment. The use of the scores from a principal component analysis to produce a process trajectory plot allows monitoring of plant progress throughout the period of operation. By observing deviations in the scores trajectory, the effect on the overall variable relationships of pilot plant experiments (such as alteration of set-points) may be examined. In this way, the pilot plant may be considered as a batch process and future work will include the application of multiway analysis techniques [26–28] in order to compare data recorded using different operating conditions.

The use of cross-correlation using latent variables rather than the raw data greatly improves the interpretative power of time delay analysis. For each response, only a few cross-correlation plots are necessary, rather than the 24 plots (i.e., one for each process variable) which traditional analysis would entail. In conjunction with information on the loadings, the latent vector cross-correlation method provides very useful information as to which process unit influences which aspect of product quality. An alternative method is to apply PLS using an X-block consisting of the predictor vector repeated in each column, but shifted back one point at a time. Using this method, the PLS regression coefficients for a one component model are found to be commensurate with the correlation coefficients calculated above.

Acknowledgements

The authors acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) of UK and the SCIENTIA Innovative Manufacturing Initiative (IMI) grant IMP/RP/01/016. They also acknowledge the support of the Centre for Process Analysis, Chemometrics and Control and the constructive discussions within the MSPC team. The suggestions of the referees are also acknowledged in their contribution to the final paper.

References

[18] W. Windig, Mixture analysis of spectral data by multivariate analysis of spectral data by multivariate...