Latent Variable Models and Big Data in the Process Industries

J. F. MacGregor, M. J. Bruwer, I. Miletic, M. Cardin, Z. Liu

ProSensus, Inc.

Ancaster, ON, Canada (Tel: 905-304-9433; email: john.macgregor@prosensus.ca)

Abstract: In the process industries Big Data has been around since the introduction of computer control systems, advanced sensors, and databases. Although process data may not really be BIG in comparison to other areas such as communications, they are often complex in structure, and the information that we wish to extract from them is often subtle.

Multivariate latent variable regression models offer many unique properties that make them well suited for the analysis of historical industrial data. These properties and use of these models are illustrated with applications to the analysis, monitoring. optimization and control of batch processes, and to the extraction of information from on-line multi-spectral images.

Keywords: Latent variables, Big Data, Process analysis, Monitoring, Optimization, Control, Batch processes, Image analysis, PLS.

1. INTRODUCTION

In the process industries work centred on Big Data has been around since the introduction of computer control systems, advanced sensors, and databases. The main emphasis in the 1970-1990 period was to collect and display data, with data analysis never being a priority. As a result, up until the early 1990's most commercial databases and historians did not allow the user to easily access and extract data that were routinely stored. Even as late as 1998 the CEO of a large data storage company asked one of the authors why a user would want to extract data from their data storehouses. In that paradigm data were largely used for short-term operator displays, then they were compressed and stored for posterity. Even if it had been possible to extract data easily, compression algorithms often rendered any retrieved data almost useless from a modelling and analysis point of view.

Today much of the Big Data discussions are still focused on the collection, storage, management and display of data. However, the extraction of useable and timely information from databases has now become a major concern of the process industries. Although industrial process data may not really be Big Data in comparison to other sectors such as communications, industrial datasets are often complex in structure, and the information that we wish to extract from them is often subtle. This information needs to be analyzed and presented in a way that is easily interpreted and that is useful to process engineers.

This paper is focussed on the analysis and use of information from historical process databases for the purposes of improved understanding, process troubleshooting and monitoring, optimizing and controlling the processes. In order to do this effectively one must understand the nature of historical data that are collected routinely from operating processes. Most statistical textbooks and courses that are used to teach the analysis of data are looking at datasets that are very different from those collected on-line on nearly all manufacturing processes. In particular, textbook authors often assume that the measured variables are all independent (often from designed experiments). In practice, the measured variables from processes typically are all highly correlated and even though a thousand variables may be measured at any one time, there are only a small number of independent events driving all of them. There are also usually many missing measurements and sampling and measurement errors in all measured variables.

To analyze such data, analytical methods must be able to handle these key characteristics. This requirement excludes most of the classic statistical regression and analysis methods commonly taught in text books and courses. At ProSensus we rely mostly on multivariate latent variable methods such as Principal Component Analysis (PCA) and Projection to Latent Structures/Partial Least Squares (PLS). We do this for many very important reasons:

1. These methods handle the highly correlated (and reduced rank) nature of the multivariate process data by projecting the data into low dimensional latent variable spaces where most of the important information lies. For industrial processes with hundreds of measured variables it is not uncommon that only 2 -6 latent variables are needed.
2. Up to 20% or more of a dataset is often missing data due to bad sensors or, more commonly, different sampling frequencies. In some cases important measurements may only be available 5% of the time (e.g., infrequent GC or PAT measurements). Latent variable methods automatically impute these

missing data by the most likely values consistent with all the available measurements.

3. There is no need to prune the number of variables down to an independent subset. The latent variable models will generate a few statistically significant orthogonal latent variables for the regression regardless of the number of measured variables. This allows one to not have to prejudge the variables to use (i.e. to let the dataset speak for itself) and to prune later if desired.
4. The PLS regression models are always unique in spite of the greatly reduced rank (highly correlated) nature of the data. This stems from the fact that PLS simultaneously models both the X and Y spaces, unlike other regression methods that model only the correlation from the X to the Y space but ignore a model of the correlation among the X space variables themselves. This feature, which is unique to Latent variable models, is responsible for most of the following special properties of PLS regression models.
5. The PCA and PLS methods we use are interpretable. Although they do not provide information on the effects of each x-variable independently (if that information is not in the data), they readily reveal how combinations of variables most relevant to variations in the y's relate to changes in operating conditions such as good versus bad operation.
6. These methods allow for process monitoring (MSPC) using only a few multivariate statistics (SPE, T^2) that summarize the health of the operating process.
7. For on-line applications such as soft sensors or inferential models the methods provide an effective assessment of the validity of incoming measurements prior to their use in the models.
8. They provide causal models in the low dimensional latent variable spaces within which one can optimize process operating conditions. To be able to use historical process data (which we often have in abundance) in this way is a huge advantage over other regression methods (e.g., multiple linear regression, neural networks) that offer no such causality.

The above properties of latent variable models makes them ideally suited for use on historical process data. By contrast, the common regression methods do not offer these advantages.

To illustrate the use of these methods in practice we consider several industrial applications, including: (i) troubleshooting process problems, (ii) monitoring (MSPC) a process; (iii) optimizing a process using historical data, and (iv) advanced control. In all cases we will focus the applications on batch processes since this is a rather novel area in process systems engineering. However, the methods are equally applicable to continuous processes.

2. MULTIVARIATE ANALYSIS OF PROCESS DATA

2.1 Analysis of Process Operations

Consider a batch process for the manufacture of a herbicide (Garcia-Munoz, et al., 2003). The objective was to troubleshoot the process and uncover what recipe variables or operating histories were related to the very high number of bad batches. The data blocks shown in Figure 1 comprise an initial condition block (Z) containing the initial chemistry and charge amount of the materials and the timing of events to be introduced in each batch, an array of process variable time histories throughout the batch (trajectories - X) and the final product attributes (Y). The data set contained approximately 200,000 values. In some cases with multiple batches in series and intermediate continuous processes, there are many more blocks and values. The ProMV latent variable software (ProSensus Inc., 2015d) is designed to handle such problems.

.

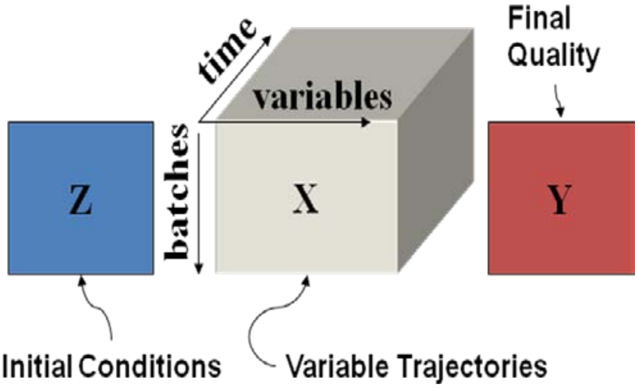


Fig. 1: Process data for batch herbicide manufacture

In spite of the large number of variables and observations, for this application the important information in the data could be summarized in terms of only two latent variables. The latent variable plot is shown in Figure 2 where each batch is summarized by two latent variable scores (good batches are shown in black and bad batches are shown in red). Note the quite clear separation of good and bad batches.

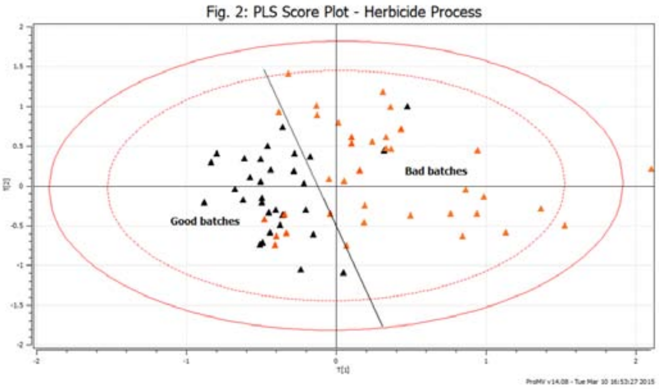


Fig. 2: PLS Score Plot – Herbicide Process

Contribution plots can then be used to show which variables contribute to the difference in these two classes – Figure 3 shows the chemistry variables contributions to the difference from good to bad batches and Figure 4 shows the contributions of the time varying process variables (trajectories) from good to bad batches. Note in the latter that the contributions of the variables are quite different at different times as one would expect in a batch process. This latter plot also serves to illustrate that the PLS batch models are in effect time varying nonlinear models. The above example illustrates how powerful and easy is the analysis of batch data by latent variable models.

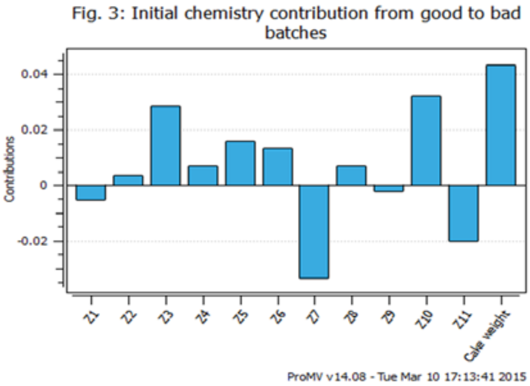


Fig. 3: Initial chemistry contribution from good to bad batches

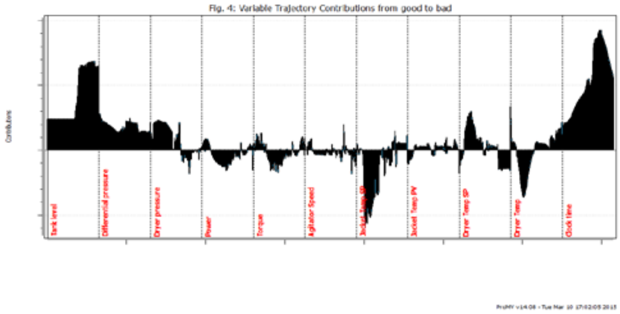


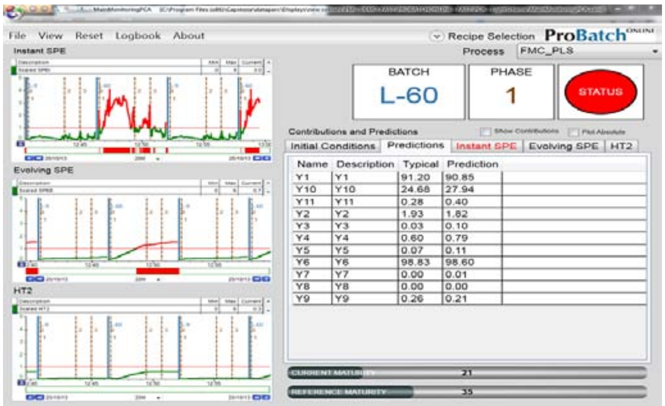
Fig. 4: Trajectory variables contributions from good to bad batches

Note that conferences impose strict page limits, so it will be better for you to prepare your initial submission in the camera ready layout so that you will have a good estimate for the paper length. Additionally, the effort required for final submission will be minimal.

2.2 *Monitoring Processes (MSPC)*

Statistical Process Control (SPC) is usually associated with univariate charts on quality attributes. But if one wishes to monitor a process that has many quality attributes and large numbers of correlated process variables, this approach is neither reasonable nor efficient. Multivariate SPC is based on monitoring any process using two multivariate statistics: (1) a summary statistic (i.e. Hotelling's T^2) for all the scores which measures movement of the process in the latent variable space (representing the magnitude of correlated variation in the variables), and (2) the squared prediction error (SPE) which is a measure of the distance that the process has moved off the latent variable plane (i.e. representing a breakdown in the correlation pattern between the variables) (Kourti and MacGregor, 1996). In this case the latent variable model is built only on data that define "common cause" variation (e.g., a representative sample of good batches in the herbicide example).

Figure 5 shows a screen shot from ProBatch Online (ProSensus Inc., 2015c) used to monitor batch processes as they evolve. The screen shows that the batch identified as L-60 is currently in Phase 1 of its operation, and is not in a state of statistical control. The instantaneous and cumulative SPE statistics in the top left and centre left (last 3 batches and the current batch are shown) have clearly exceeded their control limits although T^2 (bottom left) is still below its limit. The space in the lower right currently shows the predictions for the quality attributes at the end of the batch based on the information up to the current point. Contribution plots are also displayed there (by selecting the appropriate tab) for any statistic that is out-of-control and by clicking on any large contributors, univariate plots for those variables can be displayed. The bars at the bottom of the screen reveal that the current batch (at only 21% of maturity) is proceeding much slower than the reference batch (35% maturity) at this point of time



Name	Description	Typical	Prediction
Y1	Y1	91.20	90.85
Y10	Y10	24.68	27.94
Y11	Y11	0.28	0.40
Y2	Y2	1.93	1.82
Y3	Y3	0.03	0.10
Y4	Y4	0.60	0.79
Y5	Y5	0.07	0.11
Y6	Y6	98.83	98.60
Y7	Y7	0.00	0.01
Y8	Y8	0.00	0.00
Y9	Y9	0.26	0.21

Fig. 5: ProBatch Online screen shot for monitoring a batch process

2.3 *Process Optimization*

Latent variable models built on historical data provide causal models in the latent variables. This means that these models can be used to optimize the process as long as the optimization is done in the latent variable space (Jaeckle and MacGregor, 2000; Yacoub and MacGregor 2004). Such an optimization can be achieved by moving linear combinations of the process variables that define the scores. Unconstrained optimization with latent variable models is very simple. One need only determine the scores (τ_{des}) corresponding to a desired set of y's and then use the X-space model to obtain the solution for the optimal process conditions:

$$y=C\tau + e; \quad \tau_{des} = C(C^TC)^{-1}y_{des}\,; \quad x_{new} = P\,\tau_{des} \qquad (1)$$

For the batch herbicide optimization the Model Explore feature in ProMV can be used to achieve this by exploring the best combination of predicted y's by moving the curser in the latent variable space and then seeing the corresponding initial chemistry, timing and the batch process trajectories needed to achieve this. To incorporate constraints a formal numerical optimization in the latent variable space is needed

ProSensus has used such optimization in the latent variable space extensively not only for process optimization, but for scale-up and production transfer, and for rapidly developing new products and formulations.

2.3 Process Control

Most batch processes run with an automation layer that downloads the recipe and set-points for the desired product, sequences the charging and heating of the reactor and controls the process variables about their specified trajectories for the duration of the batch. However, almost no batch processes run with a higher-level advanced control system that optimizes the yields and controls all the properties of the final product at the end of the batch. Such a model predictive control for batch processes is quite different from continuous processes in that none of the final quality attributes or yields is measureable until the end of the batch (and usually available only hours or days later from the QC laboratory). The control/optimization problem is therefore mainly one of continually updating the prediction of the properties of the final product based on the batch initial conditions and the evolving process trajectory data. As illustrated in the batch monitoring section above, this updated prediction is very efficiently achieved using batch PLS models. Model predictive optimization/control is then achieved by taking mid-course corrective action to alter the process manipulated variable values or trajectories at one or more key decision points throughout the batch. This type of control is similar to the NASA moon missions where small rocket burns are made to alter the path of the rocket trajectory at a few key decision points during the flight. Since most batch processes currently operate in an open-loop mode and still produce reasonable product, such control at only a few decision points to compensate for any disturbances is quite sufficient (Yabuki and MacGregor, 1997; Flores-Cerillo and MacGregor 2004).

ProSensus' ProBatch Control (ProSensus Inc., 2015b) is a product based on these supervisory model predictive control concepts. It has been successfully commissioned in the food and chemical industries. Here we describe some of the results achieved over the past five years on a batch process at an international food company. The system has controlled over 800,000 batches with >99% up-time. Optimizations are performed at two mid-course decision points using different objective functions. It has achieved >50% reduction in all final quality attributes (see Fig. 6), increased productivity by 20-30% and has reduced operator involvement five-fold.

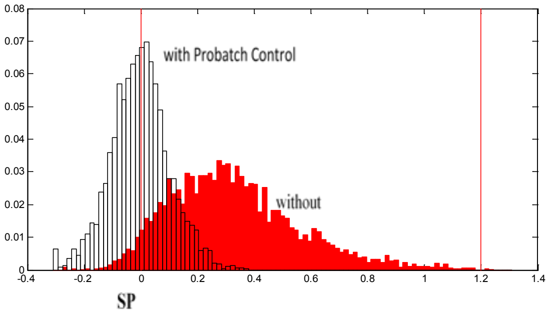


Fig. 6: Results of ProBatch Control on one quality attribute

3. MULTIVARIATE DIGITAL IMAGING FOR PROCESS MONITORING AND CONTROL

Another active area for the application of latent variable modelling is in the use of multivariate image analysis to monitor and control solid and heterogeneous slurry products where spatial variations and defects are important. Industries that produce such products have very few standard measurements that can measure quality on-line, unlike the petrochemical and chemical industries that can insert sensors into liquids and gases. However, digital imagery provides an inexpensive solution. It is a form of Big Data since even a single image can represent data from millions of pixels. Latent variable models can be used to extract subtle information from these powerful sensors that relate to multiple defects and product quality (Duchesne et. al. 2012). ProSensus has successfully applied color cameras on-line in the snack food industry to monitor product quality attributes such as seasoning and seasoning distribution, texture and defect levels (Yu et. al., 2003). Control over some of the properties using feedback from the cameras has been in use now for more than a decade.

ProSensus' Baleguard system (ProSensus Inc. 2015a) is another example of multivariate image analysis used for on-line inspection of sheet and rubber bale products. It is a six-camera colour/NIR imaging system that images all six sides of rubber bales passing through it every few seconds (see Fig. 7). Multiple defects are identified, their severity quantified and bales with too many defects rejected automatically. These systems were originally installed for such inspections, but once the severity information for the multiple defects was collected by the imaging system and entered into the database as time series, PLS models were able to relate the severity of many of the defects to conditions in the upstream process. This information led, in some cases, to a more than ten-fold reduction in the severity of certain defects.

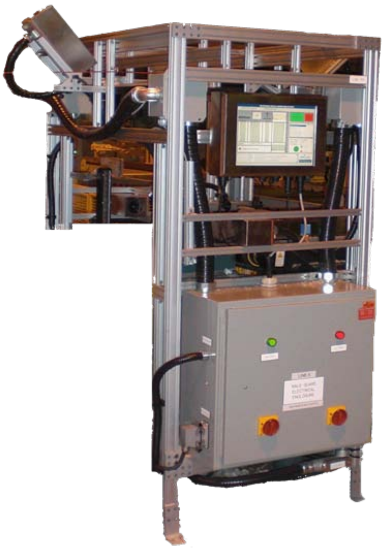
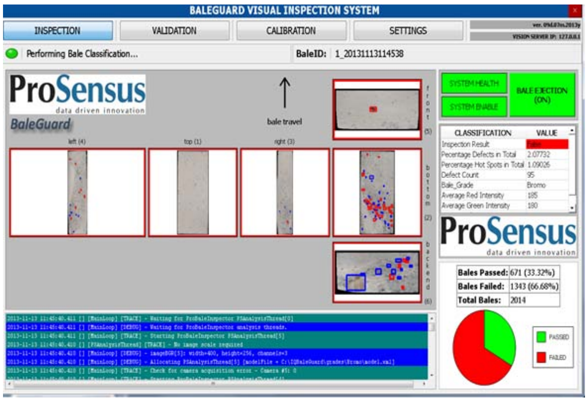


Fig. 7a: BaleGuard Hardware



VISION SERVER IP: 127.0.0.1

Performing Bale Classification...

BaleID: 1_20131113114538

CLASSIFICATION	VALUE
Inspection Result	False
Pecentage Defects in Total	2.07732
Percentage Hot Spots in Total	1.09026
Defect Count	95
Bale_Grade	Bromo
Average Red Intensity	185
Average Green Intensity	180

Bales Passed:	671 (33.32%)
Bales Failed:	1343 (66.68%)
Total Bales:	2014

```
2013-11-13 11:45:40.411 [] [MainLoop] [TRACE] - Waiting for ProBaleInspector PSAnalysisThread[0]
2013-11-13 11:45:40.411 [] [MainLoop] [DEBUG] - Waiting for ProBaleInspector analysis threads.
2013-11-13 11:45:40.411 [] [MainLoop] [TRACE] - Starting ProBaleInspector PSAnalysisThread[5]
2013-11-13 11:45:40.420 [] [PSAnalysisThread] [TRACE] - No image scale required
2013-11-13 11:45:40.420 [] [MainLoop] [DEBUG] - imageBGR[5]: width=400, height=256, channels=3
2013-11-13 11:45:40.420 [] [MainLoop] [DEBUG] - Allocating PSAnalysisThread[5] [modelFile = C:\IQBaleGuard\grades\Bromo\model.xml]
2013-11-13 11:45:40.420 [] [MainLoop] [TRACE] - Check for camera acquisition error - Camera #5: 0
```

Fig. 7b: BaleGuard Software Interface

4. CONCLUSION

In this paper we have presented what we believe to be the huge advantages of latent variable methods for extracting and using information from Big Data in the process industries. We have illustrated the use and power of these methods in the process industries for (i) the analysis and troubleshooting of process problems, (ii) process monitoring, (iii) process optimization, (iv) batch process control and (v) on-line imaging. Although we have provided very few details on each application, these are available in some of the references provided. The main message of this paper is to present the concepts of the methods and make readers aware of the scope of their application in the process industries.

5. REFERENCES

Duchesne, C., Liu, J. J., MacGregor, J. F., (2012). “Multivariate image analysis in the process industries: A review”, *Chem. & Intell. Lab. Systems*, 117, 116-128.

Garcia-Munoz, S., Kourti, T., MacGregor, J. F., Mateos A. G., Murphy G., (2003). “Troubleshooting of an industrial batch process using multivariate methods”, *Ind. Eng. Chem. Res.*, 42, 3592-3601.

Flores--Cerrillo, J., MacGregor J. F., (2004). “Control of batch product quality by trajectory manipulation using latent variable models”, *Journal of Process. Control*, 14, 539-553.

Jaeckle,C.M, and MacGregor, J.F. (2000) “Industrial Applications of Product Design Through the Inversion of Latent Variable Models”, Chemometrics & Intell. Lab. Sys., 50, 199-210.

Kourti T., MacGregor, J. F., (1996). "Multivariate Statistical Process Control methods for monitoring and diagnosing process and Product Performance", *Journal of Quality Technology*, 28, 409-428.

ProSensus Inc., (2015a). “BaleGuard Imaging System”, http://www.prosensus.ca/solutions/multivariate-machine-vision-systems/baleguard-surface-inspection.

ProSensus Inc., (2015b). “ProBatch Control: a Supervisory Model Predictive Control product for batch processes”, http://www.prosensus.ca/solutions/multivariate-analysis-software/probatch-control.

ProSensus Inc., (2015c). “ProBatch Online: a monitoring system for batch processes”, http://www.prosensus.ca/solutions/multivariate-analysis-software/probatch-online.

ProSensus Inc., (2015d). “ProMV: ProSensus multivariate analysis software”, http://www.prosensus.ca/solutions/multivariate-analysis-software/promv-desktop-multivariate-analysis

Y. Yabuki, and MacGregor, J.F. (1997) “Product Quality Control in Semi-Batch Reactors using Mid-Course Correction Policies”, Ind. Eng. Chem. Res., 36, 1268-1275.

Yacoub, F. and MacGregor, J.F. (2004) “Product optimization and control in the latent variable space of nonlinear PLS models”, Chemometrics & Intell. Lab. Syst., 70, 63-74.

Yu, H., MacGregor, J.F., Haarsma, G. and Bourg, W. (2003) “Digital imaging for on-line monitoring and control of industrial snack food processes”, Ind. & Eng. Chem. Res., 42, 3036-3044.